Binding and unbinding the auditory and visual streams in the McGurk effect

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

Binding and unbinding the auditory and visual streams in the McGurk effect

Olha Nahorna, Frédéric Berthommier & Jean-Luc Schwartz(1)

GIPSA-Lab, Speech and Cognition Department,

UMR 5216, CNRS – Grenoble University – France

Suggested running title: Binding and unbinding in the McGurk effect

Abstract

Subjects presented with a coherent auditory and visual stream generally fuse them into a single percept. This results in enhanced intelligibility in noise, or in visual modification of the auditory percept in the McGurk effect. It is classically considered that processing is done independently in the auditory and visual systems before interaction occurs at a certain representational stage, resulting in an integrated percept. However, some behavioral and neurophysiological data suggest the existence of a two-stage process. A first stage would involve *binding* together the appropriate pieces of audio and video information, before fusion per se in a second stage. Then it should be possible to design experiments leading to *unbinding*. It is shown here that if a given McGurk stimulus is preceded by an incoherent audiovisual context, the amount of McGurk effect is largely reduced. Various kinds of incoherent contexts (acoustic syllables dubbed on video sentences, or phonetic or temporal modifications of the acoustic content of a regular sequence of audiovisual syllables) can significantly reduce the McGurk effect, even when they are short (less than 4s). The data are interpreted in the framework of a-two-stage "binding and fusion" model for audiovisual speech perception.

Suggested PACS Classification numbers

Main section: 43.71

Detailed classification: 43.71.An, 43.71.Es

Keywords**:** audiovisual speech perception; multisensory coherence; conditional binding; attentional mechanisms; audiovisual fusion

I. Introduction

A. Audiovisual interactions in speech perception

It is well known that the visual modality participates in the decoding process in speech perception. Classical paradigms displaying audiovisual interaction mechanisms involve improvements in speech comprehension in noise due to lipreading (Sumby and Pollack, 1954), and the McGurk effect in which a conflicting visual input modifies the perception of an auditory input, e.g., visual /ga/ added on auditory /ba/ leading to the percept of /da/ (McGurk and MacDonald, 1976). It is generally considered that processing is done independently in the auditory and visual systems before interaction occurs at a certain representational stage, resulting in an integrated percept. A number of studies are focussed on the stage at which fusion occurs, considering late vs. early fusion in reference to categorisation. In the first case, phonetic decoding takes place independently in each sensory modality before fusion occurs, while in the second case, fusion takes place on pre-categorical representations, which means that auditory and visual representations are represented in a common format at a given stage, this format being possibly motor (Summerfield, 1987; Schwartz et al., 1998).

Other studies deal with the nature of the fusion process (e.g. Massaro, 1989). Fusion has long been considered to be automatic (Massaro, 1987; Soto-Faraco et al., 2004). However, this has been questioned in recent experiments showing that imposing high demands on the attentional system decreases the amount of audiovisual fusion (and hence the percentage of visually influenced responses to audiovisual stimuli in the McGurk paradigm), irrespective of whether the attentional load is imposed on the visual

(Tiippana et al., 2004), auditory (Alsius et al., 2005) or even the tactile system (Alsius et al., 2007).

B. A multi-stage process?

While evidence for the non-automaticity of the fusion mechanism stays compatible with a one-stage architecture, some data suggest that audiovisual interactions could intervene at various stages in the speech decoding process. A first set of data concerns the "audio-visual speech detection advantage": the presence of the speaker's face has been shown to improve the detection of speech embedded in acoustic noise (Grant and Seitz, 2000) and the temporal correlation between the auditory and visual components has been shown to play a crucial role in this process (Kim and Davis 2004). This effect occurs even in an entirely unfamiliar language, which rules out interpretation based on top-down effects and pure lipreading mechanisms (Kim and Davis 2003). The gain provided by the sight of lip movements can enable the listener to better extract acoustic cues and improve phonetic categorisation in noise (Schwartz et al., 2004). A second set of experimental data come from electrophysiological experiments displaying early latencies of audiovisual interactions in the auditory cortex (Colin et al., 2002; Besle et al., 2004; Ponton et al., 2009), suggesting that visual speech can speed up the cortical processing of the auditory input as early as 100ms after the stimulus onset (van Wassenhove et al., 2005).

This suggests that the visual speech flow could modulate ongoing auditory feature processing at various levels (Bernstein et al., 2004a; Bernstein et al., 2008a; Eskelund et al., 2011). At least, audiovisual fusion could be conceived as a two-stage process,

beginning by *binding* together the appropriate pieces of audio and video information, followed by integration per se (Berthommier, 2004). The binding stage would occur early in the audiovisual speech processing chain enabling the listener to extract and group together the adequate cues in the auditory and visual streams, exploiting coherence in the dynamics of the sound and sight of the speech input. This would result in the advantage in auditory detection of speech in noise provided by the simultaneous speaking face, which would provide temporal cues about when to listen to in the acoustic material (Grant and Seitz, 2000; Kim and Davis 2004). The gain in speech comprehension in noise due to the visual input would be partly due to this temporal cuing process (Schwartz et al., 2004). Fusion would occur at a further stage, and be more or less conditioned by this preliminary binding stage.

If binding does indeed occur prior to audiovisual fusion, then it should be possible to design experiments leading to *unbinding*. This is the objective of the present paper. The assumption is that binding and unbinding are dynamic processes, and that if a given audiovisual context provides strong evidence in favour of unbinding of the auditory and visual inputs, then a given McGurk target following this piece of context should display less audiovisual fusion, and hence less McGurk effect. This assumption is explored in two experiments in which either a coherent or an incoherent audiovisual context at various durations is presented either before a congruent “ba” audiovisual target, or before an incongruent “McGurk” target combining an audio “ba” with a visual “ga” (Fig. 1). The subject’s task is to monitor online the perception of either “ba” or “da” stimuli. The expectation is that the subject will experience less often a fused “da” percept for McGurk stimuli (and hence, produce more often a pure auditory “ba” response) when they are preceded by an incoherent context. In a first experiment the incoherence between the auditory and the visual context is large, with a regular rhythm of acoustic syllables

dubbed on a completely unpredictable sequence of video sentences. In a second experiment, the incoherence is largely reduced by using just sequences of syllables, with either phonetic or temporal incoherence in the acoustic and visual materials designing the audiovisual contexts. The data are interpreted in the framework of a two-stage "binding and fusion" model for audiovisual speech perception (Berthommier, 2004).

II. Experiment 1: Decreasing the McGurk effect by an incoherent audiovisual context

This first experiment aimed at displaying that it is indeed possible to modulate the McGurk effect, in a situation where exactly the same McGurk stimuli (combining an audio "ba" with a video "ga") were preceded by a speech audiovisual context made up of either consistent audio and video streams (i.e. the sound and sight of the speaker uttering sequences of syllables), or largely inconsistent audio and video streams coming from different speech materials produced by the same speaker.

A. Materials and Methods

1. Participants

19 French subjects without hearing or vision problems participated in the experiment (6 women and 13 men, between 22 and 27 years old, 17 right-handed and 2 left-handed). They all gave informed consent to participate in the experiment and were not aware of the purpose of the experiments.

2. Stimuli

Subjects were presented with audiovisual films consisting of an initial part called "context" followed by a second part called "target". The audiovisual context was either coherent or incoherent. The target was either a congruent audiovisual "ba" syllable, or an incongruent McGurk stimulus with an audio "ba" mounted on a video "ga" (Fig. 1). Congruent audiovisual "ba" syllables should be perceived as "ba", while incongruent McGurk stimuli should often be perceived as "da" (McGurk and MacDonald, 1976). Therefore, the focus was actually on McGurk targets, the congruent "ba" targets being presented only as controls.

All stimuli were prepared from two series of audiovisual material, a "syllable" material and a "sentence" material, produced by a French male speaker, JLS, with lips painted in blue to allow precise video analysis of lip movements (Lallouache, 1990). The "syllable" material consisted of 32 random sequences of 5, 10, 15 or 20 syllables (8 occurrences of each length) containing "pa", "ta", "va", "fa", "za", "sa", "ka", "ra", "la", "ja", "cha", "ma" or "na" (the "context"), followed in half the cases by "ba" and in the other half by "ga" (the "target"). The syllable rhythm being about 1.5 Hz, the sequence durations varied from less than 4 s to around 10 s depending on the number of syllables uttered. The speaker was instructed to respect a short silence between each syllable, which was necessary for further audio editing. The "sentence" material consisted of a "context" set of 32 sequences of free sentences (invented online by the speaker), which lasted around 4, 7, 10 or 13 s (the speaker was alerted when the time was up), finished in half the cases by uttering a "ba" and in the other half by a "ga" (the "target"). Recordings were digitised at an acoustic sampling frequency of 44.1 kHz and a video sampling frequency of 50 Hz (25

images per second with two frames per image). While the "syllable" and "sentence" materials were both used to prepare the "context" stimuli, only the "syllable" material provided the basis for preparing the "target" "ba" or "McGurk" stimuli.

In this experiment, a strong incoherence was produced by dubbing the audio content of the "syllable" context on the video content of the "sentence" context (Fig. 2). Acoustic files of the 64 films were processed to detect the onset of the last target syllable "ba" or "ga", and the corresponding image was also labelled. Incoherent context sequences were prepared by dubbing an acoustic stream from a "syllable" sequence on a video stream from a "sentence" sequence with the adequate duration (video sequences were cut at the beginning, and precisely synchronised with the acoustic streams so that the final syllable onsets were exactly synchronous). Coherent contexts were provided by the "syllable" context with perfect coherence between the audio and the video streams. Exactly the same audio material was presented in the coherent and incoherent contexts by dubbing the same audio files on two different video streams ("syllables" and "sentences"). Both coherent and incoherent contexts were cut just before the beginning of the "target" syllable.

Targets consisted of congruent audiovisual "ba" stimuli, and incongruent "McGurk" stimuli prepared by editing the acoustic files of films finishing with a "ga", and replacing the "ga" sound with a "ba" excerpt extracted from appropriate acoustic files. The "ba" sound was positioned exactly at the same temporal position as the "ga" sound, synchronisation being ensured by superposing temporal positions of the plosive burst at the onset of the target stimulus.

To be sure that no difference could exist in the target stimuli in various contexts, a fixed set of target stimuli (comprising "ba" and "McGurk" stimuli) was positioned at the end of

the coherent and incoherent context sequences. This set was extracted only from the "syllable" context (that is, corresponding in the visual stream to sequences of syllables, as in the audio stream). To ensure continuity between the end of the context stimulus and the onset of the target stimulus, a 200-ms transition stimulus (5 images without sound) was inserted between context and target (with a progressive linear shift from face to black from images 1 to 3, and a progressive linear shift from black to face from images 3 to 5). The same set of target stimuli was used for both contexts and for all context durations. Care was systematically taken to dub a context and a target from different recordings, so that the amount of inconsistency between context and target was basically the same for all stimuli, whatever the context and the target (subjects never complained that there was a perturbing discontinuity from context to target, discontinuity actually being very difficult to notice thanks to the mounting procedure described above)(2).

3. Procedure

The subject's task was to detect online "ba" or "da" syllables (syllable monitoring task), without knowing when they could occur in the sequence. McGurk stimuli were presented three times more than congruent stimuli, which served as controls. Altogether, 128 stimuli (16 congruent targets and 48 McGurk targets, in both contexts, equally divided into the four possible context durations) were presented, randomly organised in 4 32-stimuli films (blocks). Notice that each film contained a random succession of coherent and incoherent contexts at all durations (this was NOT a context-blocked experiment). All acoustic files were globally normalised in intensity to ensure that they were presented at the same level.

The experiment consisted of syllable monitoring with two possible responses – "ba" or "da" (with one button for "ba" and one for "da", the order of buttons being equally distributed across subjects). Therefore, monitoring responses could occur at any time.

All experiments were carried out in a soundproof booth with the sound presented through an earphone at a fixed level for all subjects, the level being adjusted to be comfortable for the task (around 60 dB SPL). The video stream was displayed on a screen at a rate of 25 images per second, the subject being positioned at about 50 cm from the screen. Instructions were to constantly look at the screen, and each time a "ba" or a "da" was perceived, to immediately press the corresponding button (displayed by the experimenter at the beginning of the experiment).

4. Processing of responses

The expectation in Experiment 1 was that the size of the McGurk effect, estimated by the proportion of "da" responses to the incongruent McGurk target, should be lower in the incoherent context compared to the coherent one. This modulation in the size of the McGurk effect could also be dependent on the context duration. The number of "ba" and "da" responses to the targets was computed for each subject and each condition (coherent vs. incoherent context, context duration, congruent vs. McGurk target). Since the task was syllable monitoring and the subjects did not know when the targets would occur, they could detect "ba" or "da" at any time and also fail to detect the target (failures either due to lack of response or multiple different responses to the target stimulus).

Analysis of response times enabled us to specify a protocol in which only responses within 3500 ms after the target syllable acoustic onset were considered. In the case of two different responses inside this window, the responses were discarded. Altogether (that is adding the number of lack of responses or different responses to the target), this resulted in a total of 6.4 % of cases with no response to a target stimulus. This amount is rather large, but not that surprising considering that the subjects only had two possible answers at their disposal while McGurk stimuli could result in percepts other than “ba” and “da” in French (Cathiard et al., 2001), and that they had less than 3.5 s to answer online.

Response times were defined as the time separating the plosive burst at the onset of the target stimulus and the response measured on the experimental system (experiments performed using Presentation® software (Version 0.70, www.neurobs.com). For each (subject, target, context, duration) condition, the mean response time was estimated by averaging the response times for all stimuli in the corresponding condition.

5. Statistical analyses

ANOVAs were performed on proportions of “ba” responses over the total number of “ba” plus “da” responses (ignoring cases where no response was provided by the subjects), after processing them with an asin(sqrt) transform to ensure quasi-Gaussian distribution of the variables involved. A systematic check was made that other analyses performed either on the proportions of “ba” responses over the total number of stimuli (“ba” plus “da” plus no response) or on the proportions of “da” responses over the total number of stimuli provided the same significant and non-significant effects.

It is now acknowledged that the McGurk effect is largely subject dependent (Schwartz, 2010). This may raise a problem in the present experiment, since subjects displaying smaller McGurk effects are likely to display smaller context effects. This could be dealt with by applying a transformation such as the z-score, which is likely to normalize inter-subject variability(3). Unfortunately, we checked that systematically applying this transformation on probabilities of responses did not remove much variability, and basically changed nothing in the statistical analyses.

Therefore we decided to introduce the factor "subject" as a random-effect factor in all the ANOVAs. Additionally, we systematically performed analyses with "subject" as a fixed-effect factor, to check if some effects, not significant in random-effect analyses, could appear in fixed-effect analyses (hence displaying an effect significant for the group of subjects tested in the experiment, but not strong enough to be generalisable with confidence) and hence suggest a trend for further studies. We shall mention the results of these analyses when they provide potential additional information.

Considering response times, an ANOVA was performed on the logarithm of these values for ensuring normality of the distributions, once again introducing the factor "subject" as an additional random-effect factor.

B. Results

1. Global scores

The results about subjects' responses (proportion of "ba" responses relative to the total number of "ba" + "da" responses) are set out in Fig. 3a. It appears that while "ba" targets

are indeed classified as "ba" in both contexts, McGurk targets produce a smaller proportion of "ba" responses, but this proportion is larger in the incoherent than in the coherent context. A three-factor ANOVA was performed with "target" and "context" as fixed-effect factors, and "subject" as a random-effect factor.

There is a very significant "target" effect, which displays the McGurk effect (fewer "ba" responses with the McGurk stimuli, around 60% in the coherent context, which is classical in French, see Cathiard et al., 2001, vs. close to 100% for congruent targets, $F(1,18)= 26.50$, $p<0.001$). This amount of fusion – around 40% in the coherent context – is classical in French (see Cathiard et al., 2001), and typically lower than in English (Colin and Radeau, 2003), where fusions, though highly dependent on the experimental material and conditions, can increase above 50% and up to 89% (as in the original study by McGurk and MacDonald (1976)). This is likely due to differences in the phonological systems of English and French, considering that "tha" attracts a significant amount of the fusion response in English, while it is not part of the phonological system in French.

The very significant "context" effect ($F(1,18)= 20.47$, $p<0.001$), reflecting the increase of "ba" responses with an incoherent context, is essentially due to McGurk stimuli, as shown by the significant interaction between "target" and "context" ($F(1,18)=22.84$, $p<0.001$). Post-hoc analysis confirms that the increase in the proportion of "ba" responses to McGurk targets from the coherent (60%) to the incoherent context (80%) is significant ($p<0.001$).

The "subject" effect is not significant ($F(18,17.6)=1.13$, $p=0.4$) but there is a very significant "subject"-"target" interaction ($F(18,18)=17.93$, $p<0.0001$), due to strong inter-individual differences in the McGurk effect, as shown by most studies (Schwartz, 2010). We show in Fig. 3b the comparison of scores for McGurk stimuli for each subject

and for the two contexts. The interaction between the factors “context” and “subject” is not significant ($F(18,18)=0,87$, $p=0.61$).

2. Role of duration

Focussing on the stimuli of interest that are the McGurk targets (the congruent ones being only a control), a second three-factor ANOVA was performed with the factors “subject” (random effect), “context” and “context duration” (Fig. 3c). The “subject” and “context” factors are significant, both individually (“subject” factor: $F(18,17)=24.26$, $p<0.001$; “context” factor: $F(1,18)=25.21$, $p<0.001$) and in interaction ($F(18,54)=4.80$, $p<0.001$).

The effect of “duration” is marginally significant ($F(3,54)=2.79$, $p=0.049$), though not in interaction with “context” ($F(3,54)=1.26$, $p=0.2981$) or “subject” ($F(54,54)=0.93$, $p=0.6031$). We show in Fig. 3d individual scores for each subject in the two contexts and for the four durations, showing that no clear picture emerges in terms of duration. The effect of duration is weak (and is lost in other analyses we made, either with “subject” as a fixed-factor, or with the z-score transform instead of the asin(sqrt) transform). The variation of scores with duration, showing in Fig. 3c a kind of U-curve with larger “ba” scores for the smallest and the largest duration, does not lead to any clear-cut explanation, which suggests that this is perhaps not a significant aspect of our data. Importantly, Fig. 3c shows that the “unbinding” effect due to incoherent context is already obtained with the shortest duration (5 syllables, less than 4 s), and is as large as with the largest duration.

In a further analysis[4], we tested whether the perception of a given stimulus (made of a context plus a target) could be influenced by the context of the previous stimulus. The

results are shown in Fig. 3e. It appears that a previous incoherent context tends to increase the global amount of "ba" responses by about 4% compared to a previous coherent context, whatever the context of the present sequence. This is confirmed by a three-factor ANOVA with the factors "subject" (random effect), "present context" and "previous context" (fixed effect), displaying not only a significant effect of the present context ($F(1,18)=25.79$, $p<0.001$) but also a significant effect of the previous context ($F(1,18)=7.2$, $p=0.015$), with no significant interaction between the two variables ($F(1,18)=0.02$, $p=0.88$). This result must be interpreted with some caution, since the experimental paradigm did not allow perfect control of the possible differences between targets in this analysis. Indeed, though the set of target stimuli has been controlled for being exactly the same from one context or one duration to the other, it has not been controlled for being the same from one previous context to the other. Keeping this in mind, the effect of previous context could be interpreted as showing a slow drift of the unbinding process, according to which incoherence at one time would increase the amount of unbinding later in the stimulus.

However, the effect of the previous context could also be interpreted as a decision bias, linked with "recalibration", a McGurk after effect in which fusion perception can bias further perception of auditory stimuli (Bertelson et al., 2003). Recalibration concerns mechanisms where listeners adjust their phoneme boundaries to the prevailing speech context (Vroomen & Baart, 2011). The consequence of recalibration is that if a given McGurk stimulus is perceived as "da", the next auditory "ba" stimuli, if ambiguous, are more likely to be perceived as "da". It could hence be envisioned that the next McGurk stimuli could also be influenced (though this has never been explicitly tested, to our knowledge). To attempt to disentangle binding effects from decision biases linked to recalibration, we performed separate analysis of responses to McGurk stimuli

depending on context, previous context, previous target stimulus ("ba" vs McGurk) and previous response. Notice that in this analysis, it is not possible to study each subject independently, since a number of cases are not displayed (those with previous "da" responses for subjects displaying no McGurk effect in general or in the incoherent context). Therefore, all subjects are grouped together in this analysis, no standard error is available and no ANOVA is attempted. The pattern of responses is however very clear (Fig. 3f). It can be described in three points.

1. There is a strong effect of previous response and the effect is perfectly in line with the effects described by Vroomen & Baart (2011) for decision biases from audiovisual to ambiguous auditory stimuli. Indeed, "ba" vs. "da" responses to previous McGurk stimuli produce a large bias respectively towards further "ba" vs. "da" responses to next McGurk stimuli. Such large effects (around 50%) are not uncommon in recalibration effects from audiovisual to auditory stimuli (Vroomen & Baart, 2011, Fig. 1). Conversely, previous "ba" stimuli, providing almost only "ba" responses (the very small amount of "da" responses to previous "ba" stimuli have been discarded from Fig. 3f) produce a trend towards less "ba" responses to next McGurk stimuli, compared to cases with "ba" responses to previous McGurk stimuli. This is likely due to adaptation mechanisms, where a non-ambiguous audiovisual "ba" leads to more "da" responses to next ambiguous auditory stimuli, though with smaller ranges of effects (according to Vroomen & Bart (2011, Fig. 1)).

2. Once the previous decision is taken into account, there stays no clear effect of previous context (in Fig. 3f, compare the first set of three bars with the third one, or the second one with the fourth one); hence, the contextual effect displayed in

Fig. 3e is likely to be due only to contextual decision bias rather than to binding/unbinding processes.

3. The modulating role of present context is clearly displayed in all cases of previous target/response (in Fig. 3f, compare the first set of three bars with the second one, or the third one with the fourth one). Furthermore, while this modulating role is somewhat lower for previous "ba" responses (to either McGurk or "ba" stimuli), probably because of ceiling effects, it is largely amplified for previous "da" responses to McGurk stimuli since the McGurk effect is amplified in this case.

Altogether, the amount of unbinding is already large for short incoherent context durations (5 syllables, less than 4s) and is as large as for the longest incoherent context duration (20 syllables), with possible small additional effects of duration, while the possible trend of a slow increase in unbinding for long incoherence durations seems due in fact to decision biases from one stimulus to the next.

3. Response times

Response times, defined as the time from the onset of the opening gesture for the consonant (beginning of the acoustic burst and the corresponding image) to the time of pressing the button signalling "ba" or "da" detection in the monitoring procedure, are shown separately in Fig. 4a, for each context and each target. Surprisingly, while "ba" targets seem to be processed quicker (mean value 550 ms) than McGurk ones (mean value 577 ms), there does not seem to be any effect of context. A three-factor ANOVA performed with the factors "subject" (random effect), "target" and "context" actually displays no significant effects of "target" ($F(1,18)=2.91$, $p=0.105$) or "context"

($F(1,18)=0.77$, $p=0.393$) nor of their interaction ($F(1,18)=0.1$, $p=0.758$). Interestingly, the fixed-effect analysis (with "subject" as a fixed factor) let the effect of "target" emerge significantly ($F(1,18)=6.35$, $p=0.021$). This effect should be tested for possible generalisation in further studies.

The "subject" factor is significant in isolation ($F(18,10.2)=14.85$, $p<0.001$) and close-to-significant in interaction with the "target" factor ($F(18,18)=2.18$, $p=0.054$) but the interaction between "context" and "subject" is not significant ($F(18,18)=0.70$, $p=0.77$), confirming that "context" plays no role in these data. We display in Fig. 4b the mean response times by context duration. This figure highlights the possible modulation of response times according to the target and independent of the context, while context duration does not seem to play a role here.

Therefore, there is perhaps a trend that McGurk targets are processed more slowly than the "ba" targets in both contexts. Remarkably, there is no significant difference in response times from one context to the other for both types of targets.

C. Discussion

Analysis of subjects' responses (Fig. 3) provides a rather clear picture: an incoherent audiovisual context at least 5-syllables long is enough to decrease the McGurk effect significantly. The analysis of dependence on context duration in the incoherent case (Fig. 3c) shows that the effect of context can be produced rapidly: the shortest duration (5 syllables, less than 4s) suffices to produce a large reduction in the McGurk effect, as large as the longest one (20 syllables).

A simple interpretation of these results could be that the reduction in the McGurk effect is simply due to the subject no longer looking at the video input when incoherent, and hence answering “ba” by discarding the visual input. This assumption is quite unlikely, considering the nature of the task, in which the subject is asked to perform syllable monitoring, with syllables arriving at unpredictable positions in the films, and with coherent and incoherent contexts randomly mixed in the presentation all along the films. The fact that subjects display a significant amount of McGurk effect in the incoherent context confirms that they do indeed monitor the video as well as the audio inputs.

Our interpretation is that the auditory and visual streams are bound in the listener’s brain in the default state. The incoherent context provides evidence to “unbind” them, inducing the listener to consider that the audio and video streams belong to different sources that should perhaps NOT be processed together, and hence drives the subject to reduce the role of the visual component of the McGurk target in the global response. The fact that the shortest duration produces as large an effect as the longest one shows that unbinding can be quick.

III. Experiment 2: Testing the role of phonetic vs. temporal incoherence in the McGurk modulation process

Considering that a context-driven modulation of the McGurk effect was clearly displayed in Experiment 1, Experiment 2 aimed at better understanding what kind of audiovisual incoherence was able to produce this modulation. The incoherence between the audio

and video streams in the incoherent context was very large in Experiment 1 (i.e., audio syllables vs. video sentences). In Experiment 2, we attempted to isolate simple cues enabling to produce audiovisual incoherence. We focused on two such cues, which are phonetic and temporal incoherence.

A. Materials and Methods

1. Participants

20 French subjects without hearing or vision problems participated in the experiment (5 women and 15 men, between 20 and 28 years old, 19 right-handed and 1 left-handed). They all gave informed consent to participate in the experiment, and were not aware of the purpose of the experiments.

2. Stimuli

All the stimuli in Experiment 2 were based on the "syllable" sequences, in which we produced some incoherence just by manipulation of the audio content of the context (Fig. 5). In this experiment, since context and targets all came from the same audiovisual material (the "syllable" context), we could maintain perfect continuity between context and target in the video stream, and hence no transition stimulus was necessary. Target stimuli were exactly the same in the coherent and incoherent contexts. However, they were not the same from one context duration to the other (since they were produced in different sequences).

In a first manipulation (phonetically incoherent context, P) we permuted the audio content from one syllable to the other. To maximize the chance that the audio-visual incoherence would indeed be perceivable for each context syllable, syllables were firstly organised in five groups known to be visually rather distinguishable (visemes): “pa, ma”, “fa, va”, “ta, na, sa, za”, “cha, ja” and “ka, la, ra, ga”. Then the audio content of each syllable was permuted with the content of a syllable from a different group. For each syllable**,** care was taken to maintain perfect synchrony between the sound and the image by dubbing the sound with the burst onset at exactly the same position as the original sound.

In a second manipulation (temporally incoherent context, T), we slightly advanced or delayed each audio syllable at random from 30ms audio lead to 170ms audio lag. This was aimed at staying within an “integration window” (Van Wassenhove et al., 2007) in which the McGurk effect has been shown to hardly vary. The objective was to see if random delays imposed within this window would perturb the possibility of linking the audio and video streams into a coherent context. Delays were selected at random from a fixed set of delays: [-30, 20, 70, 120, 170ms]. Notice that though precise control of the temporal delay can be done in the audio file, which is sampled at 44.1 kHz, the 50-Hz sampling in the video file makes precise control of the delay less efficient inside a 20-ms window, since applying a systematic delay of the audio file inside this 20-ms window would not be detectable. However, the incoherent delays from syllable to syllable are clearly detectable in the auditory input.

In the last incoherent context (PT), both phonetic and temporal manipulations were applied in exactly the same way as in the two previous contexts.

3. Procedure

Exactly the same procedure was used as in Experiment 1. However, this experiment, realized previously to Experiment 1[5], was implemented on a lab-developed software different from the Presentation® platform (used in Experiment 1), in which response times could unfortunately not be measured with confidence; they will accordingly not be presented here.

4. Analysis of responses

In Experiment 2, the expectation was that both the phonetic and temporal incoherence would produce a reduction in the McGurk effect, and that the detrimental contextual effect would be the combined effect of the phonetic and temporal incoherence in the last context PT. Since target stimuli were not the same from one context duration to the other, the duration effect was not tested there.

The same procedure as in Experiment 1 was used, with 168 stimuli (42 congruent targets and 126 McGurk targets, in all four contexts, equally divided into the four possible context durations) randomly organised in 4 32-stimuli films containing a random succession of coherent and incoherent contexts of all durations.

Post-processing of subjects' responses was the same as in Experiment 1: only responses within 3500 ms after the target syllable acoustic onset were considered, and in the event of two different responses inside this window, the responses were discarded. Altogether this resulted in a total of 5.4% of cases with no response to a target stimulus. All further ANOVAs were performed on proportions of "ba" responses over the total number of "ba"

plus “da” responses, these proportions being processed by an asin(sqrt) transform to ensure quasi-Gaussian distribution of the variables. A systematic check was made that other analyses performed either on the proportions of “ba” responses over the total number of stimuli (“ba” plus “da” plus no response) or on the proportions of “da” responses over the total number of stimuli provided the same significant and non-significant effects.

All statistical analyses were done in the same way as in Experiment 1.

B. Results

Subjects’ responses in Experiment 2 are displayed in Fig. 6a. They show a reduction in the McGurk effect (increase in the number of “ba” responses for McGurk targets) in the P context and to a lesser extent in the T context, and an even larger reduction in the PT context. A three-factor ANOVA was performed with “target” and “context” as fixed-effect factors, and “subject” as a random-effect factor.

Both the “context” and “target” factors and their interaction produce highly significant effects ($p<0.001$), which summarizes the difference between responses to “ba” and McGurk targets, and the selective influence of “context” on McGurk targets. Post-hoc analyses show that the P and PT conditions are both significantly different from the coherent context for McGurk stimuli ($p<0.01$), while T and coherent context on one hand, and P and PT on the other hand, do not differ significantly. The “subject” factor is not significant by itself or in interaction with the “context” factor, but the interaction between “subject” and “target” is significant ($F(19,57)=24.6$; $p<0.001$).

Focusing on the stimuli of interest that are the McGurk targets (the congruent ones being only a control), a second three-factor ANOVA was performed with the factors "subject", "phonetic incoherence" and "temporal incoherence", considering that the four contexts (coherent, P, T and PT) could be decomposed into these two factors. The results show that the three factors are significant, particularly the phonetic incoherence ($F(1,19)=31.01$, $p<0.001$) and the temporal incoherence ($F(1,19)=12.83$, $p=0.002$), with no interaction between the two ($F(1,19)=0.02$, $p=0.88$).

The "subject" effect is significant by itself ($F(19,12.2)=34.85$, $p<0.001$), as well as in interaction with "phonetic incoherence" ($F(19,19)=2.96$, $p=0.011$), but not in interaction with "temporal incoherence" ($F(19,19)=0.58$, $p=0.88$). We display in Fig. 6b individual data per subject for McGurk targets. It appears that though the inter-subject variability is very large, and some subjects display no or very small McGurk effect in any condition (e.g. subjects 1, 2, 3, 4, 15), the coherence between the effects of contexts in all subjects is striking. Indeed, the trend that phonetic incoherence and to a lesser extent temporal incoherence increase the amount of "ba" responses appears in most subjects, in spite of inter-subject variability.

Context duration did not show much effect of duration of incoherence. Since target stimuli were not the same from one context duration to another in this experiment, this could not be tested directly. However, we also performed, as in Experiment 1, an analysis of contextual effects from one stimulus to the next one. Results are displayed in Fig. 6c. They are less clear than in Experiment 1. Actually, a three-factor ANOVA with the factors "subject" (random effect), "present context" and "previous context" (fixed effect), confirms of course the significant effect of the present context ($F(3,171)=18.6$, $p<0.001$) but the effect of the previous context is not significant ($F(3,171)=2.6$,

p=0.061), while the interaction between the two variables is significant ($F(9,171)=2.44$, $p=0.012$). However, the lack of control of targets in this analysis, and the rather random aspect of data in Fig. 6c make any conclusion rather uneasy, though there is perhaps a trend that the coherent previous context produces a smaller amount of "ba" responses whatever the present context.

Once again, this could be due either to a binding/unbinding effect of context, or to a recalibration effect associated with the amount of "da" responses. Therefore, as in Experiment 1, we performed separate analysis of responses to McGurk stimuli depending on context, previous target stimulus ("ba" vs McGurk) and previous response, though the effect of previous context was not considered here because it would lead to too many different cases and non-interpretable displays. Here again (Fig. 6d), the display is very clear, with the same kind of recalibration/adaptation effects produced by the previous target and previous response, and a clear modulation by the present context. Moreover, as in Fig. 3f, the effect of context is amplified in the case of previous "da" responses to McGurk stimuli, which confirms the general trend for the effects of both phonetic and temporal incoherence displayed previously.

C. Discussion

This experiment is important for three reasons. Firstly, it confirms that context matters. Secondly, it provides a setup with no video interruption between context and target, which makes the demonstration important. Thirdly, it helps in better understanding what kind of cues seem to be exploited in the "unbinding" mechanism that we try to describe in this paper.

It appears that even rather small context incoherence seems to play a role, considering that in the phonetically incoherent context, incongruent audio and video syllables provide a rather large reduction in the McGurk effect despite their perfect temporal synchronisation. This result is a bit unexpected, considering the importance of temporal co-modulations in the speech detection paradigm (Grant and Seitz, 2000; Kim and Davis, 2004). It shows that fine analysis of the phonetic content of the audio and visual material should be part of the cues exploited for binding the audio and video streams. This result also suggests that a succession of McGurk-like stimuli would display a progressive reduction in the influence of the visual input on the subject's categorization of the audiovisual conflicting stimuli.

The fact that small random delays suffice to produce an effect (4% increase in the number of "ba" responses, small but significant) is also interesting, since we stayed within a window considered to be a plausible "integration window" for the McGurk effect (Van Wassenhove et al., 2007). This suggests that, though slight asynchrony does not impede fusion in the McGurk paradigm, accumulation of evidence for incoherent asynchronies (random from case to case) could lead to a slight perturbation in the fusion process after a few seconds. The incoherence in acoustic delays from one syllable to the next is actually audible, and could lead to various kinds of attentional effects capable of changing the efficiency of the fusion mechanism, as displayed by Alsius et al. (2005). It is likely that temporal incoherence globally decreases the amount of co-modulation in the auditory and visual channels, and hence increases evidence for unbinding, which results in a lower score of McGurk responses.

Finally, the recalibration/adaptation effects observed in Experiment 1 are confirmed in Experiment 2. These effects are seldom described in McGurk experiments and are

interesting per se. They are however not part of the focus of the present paper, and will not be considered any more in the general discussion, apart from the confirmation they provide that there does not seem to exist additional slow modulation of the binding/unbinding effects in the present data.

IV. General Discussion

A. Evidence for a binding vs. unbinding mechanism modulating fusion

Experiments 1 and 2 converge to show that McGurk fusion depends on the previous audiovisual context. This suggests that the incoherence of the audio and video streams could lead the subject to selectively decrease the role of the visual input in the fusion process. The subjects did not know when the targets would happen in the films, and the coherent and incoherent contexts were systematically mixed. This makes a simple "inattentive" mechanism - in which subjects would just drop the visual input for the task – unlikely, and rather suggests some kind of modulation of the fusion process. Our hypothesis is that modulation is driven by the output of a binding process integrating information about the coherence of the auditory and visual inputs in a given way (still to be understood). Modulation can reach about 50% of the whole McGurk effect, as displayed in Experiment 1 (from 40% to 20% of "da" responses) and in Experiment 2 (from 30% to 15% of "da" responses) with the contexts involving phonetic incoherence (P and PT).

It can appear surprising that the strong incoherence in Experiment 1 produces roughly the same amount of increase in "ba" responses for McGurk stimuli as the weaker phonetic incoherence in Experiment 2. This suggests that a close to maximum level of unbinding is already reached for a short incoherence in both cases. A difference between Experiments 1 and 2 is also the 200-ms "transition" period introduced in Experiment 1 to achieve continuity between context and target. It is known that a short temporal alert cue enhances performance in a number of tasks, such as detection of a probe inside a temporal window, or judgment of temporal order (Coull and Nobre, 1998; Correa et al., 2006). The transition cue could result in focusing the attention of the subject on the target stimulus coming just after, hence decrease the unbinding produced by the preceding context and consequently increase the amount of "da" responses. It is remarkable that this does not suffice to remove the decrease of the McGurk effect due to the incoherent context in Experiment 1. It is also important to obtain a context effect in Experiment 2 without this alert cue and with a perfect continuity between context and target.

The existence of a two-stage process has long been introduced in auditory perception through "Auditory Scene Analysis", with a first binding stage grouping together the auditory components of a given acoustic source, before categorisation processes could be applied on this source (Bregman, 1990). This paper extends this idea towards "Audiovisual Speech Scene Analysis". It is classically considered that the Auditory Scene Analysis process involves a default grouping stage followed by a possible build-up of auditory segregation (Bregman, 1990). The present data are consistent with the hypothesis of a default state of the binding mechanism in which audio and video components are fused together (leading to the McGurk effect), followed by an

"unbinding" process when evidence for different auditory and visual sources accumulates.

The McGurk effect is known to be resistant to various kinds of incongruence in the components of the sensory streams. Indeed, it can be produced despite discrepancies in the spatial localisation of the auditory vs. visual source (Bertelson et al., 1994). It is resistant to temporal asynchronies in a rather wide range, estimated to be around 200 ms, with large asymmetries from small audio leads to large audio lags (McGrath and Summerfield, 1985), and the audiovisual asynchrony, if constant, can lead to efficient recalibration processes (Vroomen et al., 2004; Navarra et al., 2005). It is even displayed with incoherence of source identity, with a female face dubbed on a male voice (Green et al., 1991; though see also Vatakis and Spence, 2007). However, the phonetic content of the auditory and visual speech material does intervene in the binding efficiency. The McGurk effect has been shown to decrease (1) when the vocalic content is conflicting between the auditory and the visual streams in addition to the consonant content (Munhall et al., 1996); (2) in case of incoherence in the time-varying aspects of speech inside the dubbed auditory and visual material (Munhall et al., 1996; Tanaka et al., 2009). This suggests that the binding process can actually display slight modulations at the level of the McGurk stimuli themselves.

Audiovisual speech binding is required for correctly associating an auditory and a visual stream in a mixture of audiovisual speech sources, e.g. in a cocktail party paradigm. It is well known that 4-month old infants are already able to correctly match a sound with a face (Kuhl and Meltzoff, 1982, 1984). When subjects are presented with an auditory source dubbed on a screen containing two faces, visual spatial attention is also able to choose between the faces when lipreading (Andersen et al., 2009). Selective attention to

the appropriate face has been correlated with steady-state visual evoked potentials on the visual scalp (Senkowski et al., 2008a). A conditional audiovisual speech binding effect has been displayed in a reverse situation in which interference from audio distractors on speechreading was shown to occur only for coherent auditory and visual streams, but not for incoherent phonetic material (Brungart and Simpson, 2005). The speech visual source has also been shown to intervene in segmentation and multistability effects (Sato et al., 2007).

B. Proposals and questions about a two-stage architecture for audiovisual fusion in speech perception

At this level, it is possible to come back to the models of audiovisual fusion available in the literature. One-stage models consider that phonetic decision operates at a given representational stage, and produces an integrated percept combining auditory and visual cues in a given way, possibly mediated by general attentional mechanisms (Fig. 7a). The present data suggest that an additional computational stage should be incorporated before decision operates (Berthommier, 2004). This involves online computation of some assessment of the coherence/incoherence of the auditory and visual inputs (C in Fig. 7b). Local coherence may also help the subject to better process the auditory and visual streams and extract adequate information for both detection and understanding of speech in noise (e.g. Grant and Seitz, 2000; Kim and Davis, 2004; Schwartz et al., 2004). Though instantaneous evidence for incoherence does not suffice to unbind the auditory and visual inputs, as displayed by the McGurk effect, accumulation of such evidence may modulate the decision process. This is displayed in

Fig. 7b by a bottom-up arrow (a). The effect of phonetic incoherence, displayed in Experiment 2, suggests that the decision stage itself could intervene in the computation of the coherence measure: this motivates the top-down arrow (b). The coherence evaluation C could result in decreasing the weight of the visual stream in the decision output if it suggests that the audio and video streams are incoherent, as happens in the various cases of incoherent contexts in Experiments 1 and 2.

Various data show that a subject is both able to perceive and estimate the discrepancy between the sight and the sound of a speaking face, and to however fuse the two inputs into a single percept (Manuel et al., 1989; Summerfield and McGrath, 1984; Soto-Faraco and Alsius, 2007, 2009). This suggests that the subject has conscious access to the output of the Coherence box, C.

The challenge for future research will be to increase our knowledge of the detailed content of this "audiovisual binding mechanism". We shall discuss successively four major questions related to the nature of the input to this binding process, the content of the dynamic system at work for achieving binding, the link between binding and decision, and the possible neuro-anatomical and neuro-physiological correlates.

1. Nature of the auditory and visual cues potentially involved in audiovisual binding

Mechanisms at work in perceptual scene analysis classically invoke the Gestaltist principle of "common fate", that is, co-evolution of the parts to group in the binding process (Bregman, 1990). In the present case, a first co-evolution cue consists of audiovisual comodulation, and particularly correlation in time between some audio

(typically global envelope or envelope in specific spectral bands) and video (typically lip or face parameter) cues. A number of papers have displayed such kinds of correlations (e.g. Munhall and Vatikiotis-Bateson, 1998; Yehia et al., 1998; Barker and Berthommier, 1999; Jiang et al., 2002; Chandrasekaran et al., 2009), and correlation in time between rms energy (particularly in the mid-to-high frequency energy-envelope) and lip area has been considered a key factor in the audiovisual speech detection advantage (Grant and Seitz, 2000; Kim and Davis, 2004).

The small but significant effect of temporal incoherence in Experiment 2 provides some evidence in favour of the role of comodulation. The effect is small, but the incoherence itself is very small also, hence the result is not so surprising. More surprising however is the strong effect of phonetic incoherence in the same experiment. Since phonetically incoherent contexts are obtained by changing the audio content while preserving as much as possible the temporal structure of the acoustic file, temporal comodulation should not be much destroyed in this context – though it is of course impossible to perfectly maintain temporal coherence while changing phonetic coherence. This shows that temporal comodulation is perhaps not sufficient to ensure binding.

The role of phonetic incoherence suggests that the fine phonetic content of each stream is determined and exploited in the binding process. In terms of cognitive architecture, this means that the audiovisual binding process receives information from auditory and visual phonetic characterization processes, as displayed in Fig. 7b. This could be compatible with the so-called “separate identification” or late-fusion process (Schwartz et al., 1998) according to which separate identification of the audio and video streams precede audiovisual fusion realizing some mixing of auditory and visual categorization processes.

Notice that the large incoherence at work in Experiment 1 obviously combines phonetic and temporal incoherence, providing altogether a large number of incoherence cues which explains the low amount of audiovisual fusion in this experiment.

A question is to know if other non-phonetic cues concerning, for example, spatial localisation, speaker identity, gender, etc, could play a role in the binding process. While some of these cues have been shown to play no or small role in the McGurk effect as recalled previously, it would be interesting to determine if longer sequences with spatial or gender discrepancies between the sound and sight could significantly modulate the McGurk effect. This raises more generally the question to know if the audiovisual binding system in speech is specific, or if it is part of a general audiovisual scene analysis mechanism.

2. Dynamics of audiovisual binding

The present work displays contextual effects in audiovisual fusion, which implies that the audiovisual fusion process is globally a dynamic system in which decision at one time depends not only on the present sensory input but also on what happened to the system previously. Better understanding of the functioning of this dynamic system involves at least four questions.

Firstly, is there a "default" state, in which the system is most likely to be, "before" any contextual influences? Of course, it could be claimed that such a situation never occurs and that the system is always in a state resulting from a history of recent influences. Since audiovisual coherence is the most likely situation, the present state should generally be considered as corresponding to a situation of coherence: hence it would be

a “bound” state. It could be also conceived that there is no default state, but a “prior” (in Bayesian terms, that is an a priori on the typical situation) in which the auditory and visual streams are supposed to be coherent, and hence should be bound together. A coherent default state or a coherent prior both converge on the idea that a classical McGurk situation should lead to fusion, which actually seems to be the case. It is also compatible with a general “compatibility bias” displayed in various experiments dealing with the fusion of conflicting cues (e.g. Yu et al., 2009; Noppeney et al., 2010). Indeed, in these studies the subjects seem to suppose at the beginning of the task that the various cues are not conflicting (this is what the authors call the “compatibility bias”), before evidence of conflict progressively lead the subjects to select one cue rather than the other.

Fluctuations could occur around this default state, even with no recent incoherent context. One question is to know if inter-individual variations of the McGurk effect could be associated with such fluctuations, asking whether subjects who do not display much McGurk effect are actually in an “unbound” state at the time of the stimulation. This is however unlikely, considering that in our data, some subjects do not experience much McGurk even after long periods of coherence in the “coherent” context (e.g. Subjects 9, 11, 13, 14, 16, 18 in Fig. 3b). Notice finally that the assumption of a default “bound” state does not mean that subjects necessarily perceive “da” in this default state. Indeed, the percept results from a fusion of available auditory and visual evidence that could each lead to various kinds of responses (such as “ba”, “da”, “ga”, “tha”, “bda”, “bga”, etc), and there is no reason that “da” should systematically win. This actually depends on many factors, including the nature of the phonological system for the subject, inter-individual variations of many kinds (see Schwartz, 2010), etc.

The second question concerns the dynamics of the unbinding process. The data in both Experiments 1 and 2 converge to suggest that unbinding is rapid and seems to display a plateau after not more than a few seconds (less than 4 seconds in our data) considering that the level of fusion in Experiment 1 is similar for all durations of context incoherence. It remains to be known what happens in this 4-s window, and we are currently exploring the role of coherent vs. incoherent contexts of smaller durations. Considering context effects from one stimulus to the next one (e.g. Fig. 3e), they seem to be linked with decision biases associated with recalibration processes rather than with binding per se.

The third question concerns the reverse binding process, particularly in an unbound configuration. Here again, binding seems to be rapid, since there is no effect of duration in the coherent context in Experiment 1, nor contextual effects from one stimulus to the other in both experiments. Shorter coherence durations will also be explored in further experiments. Another interesting point is to know what kind of information is able to reset the system and put it back in its supposedly bound default state. Could a period of stability (e.g. static faces in silence, or black screen) directly produce reset and let the system fuse again after a period of incoherence? Could a specific event, such as a flash or a clap, produce such a resetting process? This is important in the context of Experiment 1, in which there is a 200-ms fading component that is likely to decrease the role of incoherence. We have attempted to minimize these effects by applying a short fading component and carefully controlling its content so that it was as little perturbing as possible. But we cannot dismiss the assumption that incoherence effects in Experiment 1 could be *underestimated* because of a possible resetting effect due to fading (which makes these data all the more convincing).

A last point concerns the potential role of the speech motor system in the binding process (Sato et al., 2006; Schwartz et al., 2010a), in connection with recent proposals on the specific role of the motor system in audiovisual speech perception (van Wassenhove et al., 2005; Skipper et al., 2007). Motor representations have been proposed as a possible framework for audiovisual fusion (Summerfield, 1987; Schwartz et al., 1998). A multisensory-motor connection, implemented in the brain in a possible "mirror neuron system", could enable integration of the auditory and visual inputs into a coherent flow of information possibly providing a system for predicting future sensory events. This is described in computational terms in recent proposals about a "predictive coding" account of sensory integration (e.g. Kilner et al., 2007).

3. Relationships between binding and decision

If we admit, according to Fig. 7b, the existence of a two-stage architecture, the decision system has to fuse an auditory and a visual stream in a conditional way, taking into account the output of the hypothetical binding system.

A number of recent works about multisensory cue fusion claim that behavioural data nicely fit a statistically optimal Bayesian cue integration model (e.g. Ernst and Banks, 2002, for visuo-haptic fusion; Alais and Burr, 2004, for audio-visual fusion in localisation; Angelaki et al., 2010, for visuo-vestibular fusion in heading perception). Statistically-optimal cue integration is based on a probabilistic process mixing cues weighted by factors inversely proportional to their variance, thus providing more weight to a cue with low variance, and hence high "trustability". It is shown that this model is optimal in the sense that it produces a multisensory estimate with the lowest

possible variance. The Fuzzy-Logical Model of Perception, computing a product of unisensory evidence (or probabilities) and hence implicitly decreasing the role of ambiguous inputs is also classically described as a kind of statistically optimal fusion model (Massaro, 1987; 1989).

The present data show that for the same sensory inputs (same McGurk target) the output may differ because of context. A possible assumption is that the decision process comprises a weighting component driven by the output of the binding process: assuming that the audio component is the basis of the decision process, the video component would receive a lower weight in fusion if evidence for unbinding is high. Various kinds of weighted fusion models have already been proposed in the literature, including, for example, weight dependent on noise (Teissier et al., 1999, Berthommier, 2001), subject (Schwartz, 2010) and attention (Schwartz et al., 2010b). The output of the binding process may also be directly incorporated inside the decision process in a Bayesian framework, letting decision depend on both individual cues, and evidence favouring their coherence or incoherence (e.g. Yu et al., 2009, Noppeney et al., 2010).

Concerning reaction times, it is classically found that they are shorter for coherent than for incoherent stimuli in various kinds of tasks (e.g. Gondon et al., 2005), including McGurk stimuli. This is generally related to the fact that the stimuli are ambiguous: the more ambiguous they are, the longer the reaction times (Massaro and Cohen, 1983). There is a trend in Experiment 1 that responses to “ba” targets could be indeed slightly quicker than for McGurk stimuli, but surprisingly, context provides no effect (see Fig. 4a), while it plays a clear role on response probabilities (see Fig. 3a). This suggests that reaction times could depend not only on the decision process per se, but also on the evaluation done by the subject of the instantaneous discrepancy between the two

streams to integrate (see the elaboration of a "dual route" by Arnal et al., 2009, which will be presented in the next section).

4. Possible neuroanatomical and neurophysiological correlates

We shall now address the potential neuroanatomical and neurophysiological correlates of the audiovisual binding system if it exists. This system should provide a computation of some characterisation of audiovisual coherence and hence its neuronal activity should be somehow related to the amount of coherence between the auditory and the visual streams. Furthermore, it should provide a device enabling the visual input to modulate activity in the auditory regions, thus leading to effects such as the audiovisual speech detection advantage (Grant and Seitz, 2000; Kim and Davis, 2003, 2004) and early visual effects in the auditory cortex (e.g. van Wassenhove et al., 2005).

It is increasingly acknowledged that cross-modal influences intervene at the level of primary sensory cortices that were previously supposed to be sensory specific (Driver and Noesselt, 2008). Cross-modal influences in audiovisual speech perception have been displayed in functional neuroimagery (fMRI) data on both the auditory and the visual primary cortex (Calvert et al., 1997, 1999), and we have seen that these influences can occur quite early during perceptual processes (Besle et al., 2004; Colin et al., 2002). While both precortical bottom-up influences (from the superior colliculus or thalamic relays), horizontal links (directly connecting sensory cortices) and heteromodal feedback could be envisioned, the role of the heteromodal associative cortex in the Superior Temporal Sulcus (STS) is mostly cited as important in this process (Calvert et al, 2000; Ghazanfar and Schroeder, 2006). The supposedly partially speech-specific

nature of the binding process displayed in some data about the audiovisual speech detection advantage (e.g. Bernstein et al., 2004b; Schwartz et al., 2004) and the role of phonetic incoherence displayed in Experiment 2, probably discard precortical bottom-up connections as the sole or major site for audiovisual speech binding.

The role of STS (and more precisely posterior STS) is considered crucial, particularly for processing correlations between auditory and visual stimuli, which should be a basic ingredient in audiovisual speech binding (Campbell, 2008). This is in line with a recent study by Arnal et al. (2009) combining electrophysiology (MEG) and neuroimagery (fMRI). In this study, the authors suggested that there could be two separate neural routes connecting the auditory and the visual cortex. A fast corticocortical pathway, not sensitive to audiovisual incongruence, would directly connect the visual motion parameters in the auditory cortex and enable short-term predictions and modulations of the auditory activity. A slower connection would lead to the STS as a centre for estimating the degree of incoherence between the auditory and visual inputs. Feedback messages would then be sent back from the STS to the auditory and visual cortices. Interestingly, this slower route, involving the degree of audiovisual incoherence, would result in modifications of the neural response over time. This could possibly provide a correlate of the slower response times observed to McGurk than to “ba” targets in Experiment 1. Finally, a recent electrophysiological study (using MEG) by Keil et al. (2011) attempted to relate the role of temporal fluctuations in brain activity, with variability in the McGurk effect. The study correlates the amount of perceptual fusion with the state of connection between the left superior temporal gyrus and a distributed network of frontal and temporal regions. This provides interesting evidence that the McGurk effect could be a dynamic process related to the state of specific cortical areas (proposed to be located in the temporal cortex in this study).

Going towards the parietal cortex, the Supra-Marginal Gyrus (SMG) has also been proposed by Bernstein et al. (2008b) as a possible site for analysis of audiovisual incongruities. In their study, the authors presented audiovisual speech stimuli with various levels of incongruity between the audio and visual streams, and their fMRI data show that the only cortical region that demonstrated differential sensitivity to incongruity level was a subarea of the SMG.

Going up and front in the cortex, the so-called dorsal route connecting sensory and motor areas (Hickok and Poeppel, 2004) has been shown to play a role both in auditory speech organisation (Sato et al., 2004; Kondo and Kashino, 2007) and in audiovisual speech perception (Intra-Parietal Sulcus and Inferior Frontal Gyrus (Miller and d'Esposito, 2005); posterior Planum Temporale Spt; dorsal and ventral Premotor Cortex (Okada and Hickok, 2009)), particularly for binding incongruent stimuli (Jones and Callan, 2003). The dorsal route is proposed by Campbell (2008) as the natural site for dealing with correlated audiovisual speech. In the Perception-for-Action-Control Theory (PACT, Schwartz et al., 2010a), the assumption that the dorsal route could play a role in audiovisual speech binding is raised, which is also compatible with the predictive role associated to this circuit by Kilner et al. (2007), and its application to audiovisual speech processing by Skipper et al. (2007). Finally, it is not without interest to mention a recent study by Noppeney et al. (2010) on the processing of incongruent auditory and visual information on non-speech stimuli (sound and sight of actions on tools or musical instruments). In this paper, combining psychophysics and fMRI, the authors study the way incongruent cues are accumulated over time, enabling subjects to progressively shift from an initial state influenced by the "compatibility bias" mentioned previously, to an "unbound" state in which they discard incongruent and hence irrelevant information. This paradigm is quite relevant in relation with the present assumption of a binding

system assessing the amount of coherence between the auditory and visual streams and modulating the decision process accordingly (Fig. 7b). According to fMRI data in this study, the left inferior frontal sulcus (IFS) showed an "audiovisual-accumulator" profile consistent with the observed reaction time pattern; furthermore, the IFS inhibited superior temporal activations in the auditory cortex for unreliable auditory input. Altogether, Noppeney et al. (2010) conclude that: "to form decisions that guide behavioral responses, the IFS may accumulate audiovisual evidence by dynamically weighting its connectivity to auditory and visual regions according to sensory reliability and decisional relevance". Even though the task is not exactly similar, the correspondence with the architecture we proposed in Fig. 7b is quite suggestive.

Altogether, both direct cross-modal links between sensory cortices, feedback from STS, and parieto-frontal attentional modulation associated with perceptuo-motor processes, could jointly play a role in fusion (Senkowski et al., 2008b). The present experiments could provide a new paradigm for exploring the content of the binding process, by studying what kind of incoherence in the auditory and visual streams could produce unbinding assessed by McGurk fusion decrease. It is expected that "coherence" vs. "incoherence" should be fractionated into various cues possibly related to specific elements of a global cortical architecture.

V. Conclusion

The present paper reports two experiments showing that it is possible to significantly modulate the amount of audiovisual fusion in speech perception – estimated here by the degree of McGurk effect – by applying a previous incoherent vs. coherent audiovisual

context. Short context durations (less than 4s) suffice to strongly decrease the amount of fusion, and even pure phonetic incoherence keeping a high degree of temporal correlation between auditory and visual fluctuations largely diminishes the McGurk effect. A two-stage architecture with an audiovisual binding mechanism preceding speech decoding seems compatible with our experimental data. The questions for the future will be to better characterize this binding system in terms of input data, unbinding and binding dynamics, relation with the decision process, and neurocognitive architecture in the human brain.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through funding for the MULTISTAP project (MULTISTability and binding in Audition and sPeech: ANR-08-BLAN-0167 MULTISTAP).

Endnotes

(1) Corresponding author (jean-luc.schwartz@gipsa-lab.grenoble-inp.fr)

(2) Examples of stimuli for Experiments 1 and 2 are available at http://www.gipsa-lab.grenoble-inp.fr/~jean-luc.schwartz/fichiers_public_JLS/AV_Binding_demo/AV_Binding_Demo.html (date last viewed 05/10/2011)

(3) We thank one of the reviewers for having provided this interesting suggestion.

(4) Once again, we thank one of the reviewers for having provided this helpful suggestion.

(5) This experiment actually followed a first experiment providing the initial demonstration of the unbinding effect, presented in Nahorna et al. (2010). But problems linked to the control of stimuli led us introduce a new experiment with perfect control, which is described here as Experiment 1 but was performed after Experiment 2.

References

Alais, D., & Burr, D. (**2014**). "The ventriloquist effect results from near-optimal bimodal integration," Current Biology **14**, 257-62.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S.S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Current Biology **15**, 839–843.

Alsius, A., Navarra, J., & Soto-Faraco, S. S. (**2007**). "Attention to touch weakens audiovisual speech integration," Experimental Brain Research **183**, 399–404.

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (**2009**). "The role of visual spatial attention in audiovisual speech perception," Speech Communication **51**, 184–193.

Angelaki, D.E., Gu, Y., & Deangelis, G.C. (**2010**). "Visual and vestibular cue integration for heading perception in extrastriate visual cortex," Journal of Physiology **589**, 825-33.

Arnal, L.H., Morillon, B., Kell, C.A., & Giraud, A.L. (**2009**). "Dual neural routing of visual facilitation in speech processing," Journal of Neuroscience **29**, 13445-53.

Barker, J. P., & Berthommier, F. (1999). "Evidence of correlation between acoustic and visual features of speech," in Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS '99), pp. 199-202. San Francisco: USA.

Bernstein, L. E., Auer, E. T., & Moore, J. K. (**2004a**). "Audiovisual speech binding: convergence or association?," in G.A. Calvert, C. Spence C, & B.E. Stein (eds.) *The handbook of multisensory processes* (pp 203–224). Cambridge: The MIT Press.

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (**2008a**). "Spatiotemporal dynamics of audiovisual speech processing," NeuroImage **39**, 423–435.

Bernstein, L.E., Lu, Z.L., & Jiang, J. (**2008b**). "Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing," Brain Research **1242**, 172–184.

Bernstein, L. E., Takayanagi, S., Auer, E. T. (**2004b**). "Auditory speech detection in noise enhanced by lipreading," Speech Communication **44**, 5-18.

Bertelson, P., Vroomen, J., De Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: a McGurk aftereffect," Psychological Science **14**, 592–597.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (**1994**). "Exploring the relation between McGurk interference and ventriloquism," in *Proc. ICSLP 94* (Vol. 2, pp. 559–562). Yokohama: Acoustical Society of Japan.

Berthommier, F. (**2001**). "Audio-visual recognition of spectrally reduced speech," in Proceedings AVSP'01, Aalborg, pp. 183-188.

Berthommier, F. (**2004**). "A phonetically neutral model of the low-level audiovisual interaction," Speech Communication **44**, 31-41.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (**2004**). "Bimodal speech: early suppressive visual effects in human auditory cortex," European Journal of Neuroscience **20**, 2225-2234.

Bregman, A. S. (**1990**). *Auditory scene analysis*. MIT Press: Cambridge, MA.

Brungart, D.S., & Simpson, B.D. (**2005**). "Interference from audio distracters during speechreading," Journal of the Acoustical Society of America **118**, 3889-3902.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., & Mcguire, P.K. (**1997**). "Activation of auditory cortex during silent lipreading," Science **276**, 593-596.

Calvert, G.A., Brammer, M., Bullmore, E., Campbell, R., Iversen, S.D., & David, A. (**1999**). "Response amplification in sensory-specific cortices during crossmodal binding," Neuroreport **10**, 2619-2623.

Calvert, G.A., Campbell, R., & Brammer, M. (**2000**). "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex," Current Biology **10**, 649–657.

Campbell, R. (**2008**). "The processing of audio-visual speech: empirical and neural bases," Philosophical Transactions of the Royal Society of London B Biological Science **363**, 1001-1010.

Cathiard, M.A., Schwartz, J.L., & Abry, C. (**2001**). "Asking a naive question about the McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?," *Proceedings AVSP-2001*, 138-142.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. (**2009**). "The natural statistics of audiovisual speech," PLOS Comput. Biol. **5**, e1000436.

Colin, C., & Radeau, M. (**2003**). "Les illusions McGurk dans la parole : 25 ans de recherche (The McGurk illusions in speech: 25 years of research)," l'Année Psychologique **104**, 497-542.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (**2002**). "Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory," Clinical Neurophysiology **113**, 495–506.

Correa, A., Lupiáñez, J., Madrid, E., & Tudela, P. (**2006**). Temporal attention enhances early visual processing: A review and new evidence from event-related potentials," Brain Research **1076**, 116-128.

Coull, J.T., & Nobre, A.C. (**1998**). Where and when to pay attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI," Journal of Neuroscience **18**, 7426–7435.

Driver, J., & Noesselt, T. (**2008**). "Multisensory interplay reveals crossmodal influences on 'sensory specific' brain regions, neural responses, and judgments," Neuron **57**, 11-23.

Ernst, M.O., & Banks, M.S. (**2002**). "Humans integrate visual and haptic information in a statistically optimal fashion," Nature **415**, 429–433.

Eskelund, K., Tuomainen, J., & Andersen, T. S. (**2011**). Multistage audiovisual integration of speech: dissociating identification and detection," Experimental Brain Research **208**, 447-57.

Ghazanfar, A.A., & Schroeder, C.E. (**2006**). "Is neocortex essentially multisensory?,"Trends in Cognitive Science **10**, 278-285.

Gondan, M., Niederhaus, B., Rösler, F., & Röder, B. (**2005**). "Multisensory processing in the redundant-target effect: a behavioral and event-related potential study," Perception & Psychophysics **67**, 713-26.

Grant, K. W., & Seitz, P. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," Journal of the Acoustical Society of America **108**, 1197–1208.

Green, K., Kuhl, P., Meltzoff, A., & Stevens, E. (**1991**). "Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect," Perception and Psychophysics **50**, 524-536.

Hickok, G., & Poeppel, D. (**2004**). "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language," Cognition **92**, 67-99.

Jiang, J., Alwan, A., Keating, P.A., Auer, E.T., & Bernstein, L.E. (**2002**). "On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics," Eurasip Journal on Advances in Signal Processing **11**, 1174-1188.

Jones, J.A., & Callan, D.E. (**2003**). "Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect," Neuroreport **14**, 1129-33.

Keil, J., Müller, N., Ihssen, N., & Weisz, N. (**2011**). "On the Variability of the McGurk Effect: Audiovisual Integration depends on pre-stimulus Brain States," Cerebral Cortex, doi: 10.1093/cercor/bhr125.

Kilner, J.M., Friston, K.J., & Frith, C.D. (**2007**). "Predictive coding: an account of the mirror neuron system," Cognitive Processes **8**, 159-166.

Kim, J., & Davis, C. (**2003**). "Hearing foreign voices: does knowing what is said affect masked visual speech detection?," Perception **32**, 111–120.

Kim, J., & Davis, C. (**2004**). "Investigating the audio-visual detection advantage," Speech Communication **44**, 19-30.

Kondo, H.M., & Kashino, M. (2007). "Neural mechanisms of auditory awareness underlying verbal transformations," Neuroimage **36**, 123-130.

Kuhl, P.K., & Meltzoff, A.N. (1982). "The bimodal development of speech in infancy," Science **218**, 1138–1141.

Kuhl, P.K., & Meltzoff, A.N. (1984). "The intermodal representation of speech in infants," Infant Behavior and Development **7**, 361–381.

Lallouache, M.T. (1990). « Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. *(A "face-speech" workstation. Acquisition and processing of labial contours)*," *Proceedings XVIII Journées d'Etudes sur la Parole (*pp. 282-286), Montréal.

Manuel, S., Repp, B. H., Liberman, A. M., & Studdert-Kennedy, M. (1989). "Exploring the "McGurk effect"," *Paper presented at the 24th meeting of the Psychonomic Society*, San Diego.

Massaro, D. W. (1989). "Multiple Book Review of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*," Behavioral and Brain Sciences **12**, 741–794.

Massaro, D. W. (1987). "*Speech perception by ear and eye*" (320 p.), Hillsdale: LEA.

Massaro, D. W., & Cohen, M. M. (1983). "Evaluation and Integration of Visual and Auditorial Information in Speech Perception," Journal of Experimental Psychology: Human Perception and Performance **9**, 753-771.

McGrath, M., & Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," Journal of the Acoustical Society of America **77**, 678–685.

McGurk, H., & MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **265**, 746–748.

Miller, L.M., & D'Esposito, M. (**2005**). "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech," Journal of Neuroscience **25**, 5884 –5893.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (**1996**). "Temporal constraints on the McGurk Effect," Perception and Psychophysics **58**, 351–362.

Munhall, K.G., & Vatikiotis-Bateson, E (**1998**). "The moving face during speech communication," in R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by Eye II* (pp. 123–139). Sussex: Taylor and Francis.

Nahorna, O., Berthommier, F., & Schwartz, J. (**2010**). "Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context," in Proc. *AVSP2010 - International Conference on Auditory-Visual Speech Processing* (p. 150*).* Hakone, Kanagawa, Japan.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (**2005**). "Exposure to asynchronous audiovisual speech increases the temporal window for audiovisual integration of non-speech stimuli," Cognitive Brain Research **25**, 499–507.

Noppeney, U., Ostwald, D., & Werner, S. (**2010**). "Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex," Journal of Neuroscience **30**, 7434-46.

Okada, K., & Hickok, G. (**2009**). "Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data," Neuroscience Letters **452**, 219-23.

Ponton, C. W., Bernstein, L. E., & Auer, E. T. (**2009**). "Mismatch Negativity with Visual-only and Audiovisual Speech," Brain Topography **21**, 207–215.

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J-L., Cathiard, M-A., Segebarth, C. & Abry, C. (**2004**). "Multistable perception of speech forms in working memory: An fMRI study of the verbal transformation effect," Neuroimage **23**, 1143-1151.

Sato, M., Basirat, A., & Schwartz, J.L. (**2007**). "Visual contribution to the multistable perception of speech," Perception and Psychophysics **69**, 1360-1372.

Sato, M., Schwartz, J.L. Cathiard, M. A., Abry, C., & Loevenbruck, H. (**2006**). "Multistable syllables as enacted percepts: a source of an asymmetric bias in the verbal transformation effect," Perception & Psychophysics **68**, 58-474.

Schwartz, J.L. (**2010**). "A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent," Journal of the Acoustical Society of America **127**, 1584-1594.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (**2010a**). "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," Journal of Neurolinguistics xxx (2011) 1–19.

Schwartz, J.L., Berthommier, F., & Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," Cognition **93**, B69–B78.

Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (**1998**). "Ten years after Summerfield ... a taxonomy of models for audiovisual fusion in speech perception," in R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by Eye, II. Perspectives and directions in research on audiovisual aspects of language processing* (pp. 85-108). Hove (UK): Psychology Press.

Schwartz, J.L., Tiippana, K., & Andersen, T. (**2010b**). "Disentangling unisensory from fusion effects in the attentional modulation of McGurk effects: a Bayesian modeling study suggests that fusion is attention-dependent," in Proceedings AVSP2010 (pp. 23-27). Tokyo, Japan.

Senkowski, D., Saint-Amour, D., Gruber, T., & Foxe, J.J. (**2008a**). "Look who's talking: The deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions," Neuroimage **43**, 379-387.

Senkowski, D., Schneider, T.R., Foxe, J.J., & Engel, A.K. (**2008b**). "Crossmodal binding through neural coherence: implications for multisensory processing," Trends in Neurosciences **31**, 401-409.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (**2007**). "Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception," Cerebral Cortex **17**, 2387–2399.

Soto-Faraco, S., & Alsius, A. (**2007**). "Conscious access to the unisensory components of a cross-modal illusion," Neuroreport **18**, 347–50.

Soto-Faraco, S., & Alsius, A. (**2009**). "Deconstructing the McGurk-MacDonald illusion," Journal of Experimental Psychology: Human perception and performance **35**, 580-7.

Soto-Faraco, S., Navarra, J., & Alsius, A. (**2004**). "Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task," Cognition **92**, B13–B23.

Sumby ,W. H, & Pollack, I. (**1954**). Visual contribution to speech intelligibility in noise," Journal of the Acoustical Society of America **26**, 212–215.

Summerfield, Q. (**1987**). "Some preliminaries to a comprehensive account of audio-visual speech perception," in B. Dodd & R. Campbell (eds.) *Hearing by Eye: The Psychology of Lipreading* (pp. 3–51) New York (NY): Lawrence Erlbaum Associates.

Summerfield, Q., & McGrath, M. (**1984**). "Detection and resolution of audio-visual incompatibility in the perception of vowel," Quarterly Journal of Experimental Psychology **36A**, 51-74.

Tanaka, A., Sakamoto, S., Tsumura, K., & Suzuki, Y. (**2009**). "Visual speech improves the intelligibility of time-expanded auditory speech," NeuroReport **20**, 473–477.

Teissier, P., Robert-Ribes, J., Schwartz, J.L., & Guérin-Dugué, A. (**1999**). "Comparing models for audiovisual fusion in a noisy-vowel recognition task," IEEE Trans. Speech and Audio Processing **7**, 629-642.

Tiippana, K., Andersen, T.S., & Sams, M. (**2004**). "Visual attention modulates audiovisual speech perception," European Journal of Cognitive Psychology **16**, 457–472.

Vatakis, A., & Spence, C. (**2007**). "Crossmodal binding: Evaluating the 'unity assumption' using audiovisual speech stimuli," Perception and Psychophysics **69**, 744-756.

Vroomen, J. & Baart, M. (**2011**). "Phonetic recalibration in audiovisual speech," in M. M. Murray & M. T. Wallace (eds.) *Frontiers in the neural basis of multisensory processes* (pp. 363-379) Routledge: Taylor & Francis.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (**2004**). "Recalibration of temporal order perception by exposure to audio-visual asynchrony," Cognitive Brain Research **22**, 32–35.

Van Wassenhove, V., Grant, K.W., & Poeppel, D. (**2005**). "Visual speech speeds up the neural processing of auditory speech," Proceedings of the National Academy of Sciences **102**, 1181–1186.

Van Wassenhove, V., Grant, K.W., & Poeppel, D. (**2007**). "Temporal window of integration in bimodal speech," Neuropsychologia **45**, 598-607.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (**1998**). "Quantitative association of vocal-tract and facial behavior," Speech Communication **26**, 23–43.

Yu, A. J., Dayan, P., & Cohen, J. D. (**2009**). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. Journal of Experimental Psychology: Human Perception and Performance **35**, 700-717.

Figure captions

Figure 1 - Experimental paradigm

Two contextual audiovisual stimuli (either coherent or not) precede two target audiovisual stimuli (a congruent audiovisual " ba " or a McGurk stimulus combining an audio " ba " with a visual " ga ").

Figure 2 - Stimuli setting for Experiment 1

The coherent context consists of a sequence of 5, 10, 15 or 20 syllables. In the incoherent context, the auditory content is the same, but the visual content is replaced by a series of sentences matched in global duration. A 200-ms stimulus allows the transition between context and target. The target is either a congruent audiovisual " ba " or a McGurk stimulus combining an audio " ba " with a visual " ga ".

Figure 3 - Results of Experiment 1: percentages of "ba" responses

(a) Percentage of " ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (left) vs. incoherent (right) contexts. Error bars display standard errors computed from the residual error in the three-factor ANOVA (subject variability removed).

(b) Individual results for McGurk targets. For each of the 19 subjects, the percentage of "ba" responses for the McGurk target is displayed in the coherent and incoherent contexts.

(c) Percentage of " ba " responses for " McGurk " stimuli, in the coherent (left) vs. incoherent (right) contexts for the four context durations.

(d) Individual results for McGurk targets, in the coherent (left) and incoherent (right) contexts. For each of the 4 durations, the percentage of "ba" responses for the McGurk target is displayed for the 19 subjects.

(e) Effect of the precedent context, depending on the context of the present stimulus. Percentage of " ba " responses for McGurk stimuli in a coherent (left) or incoherent (right) context whether the precedent context is coherent (in dark grey) or incoherent (in light grey).

(f) Analysis of responses to McGurk stimuli depending on context ("Coh" for coherent, "Incoh" for incoherent), precedent context ("Prec coh" for coherent precedent context, "Prec incoh" for incoherent precedent context), precedent target stimulus ("Prec Ba" vs "Prec McGurk") and previous answer ("Ans ba" for previous "Ba" target, "Ans ba" and "Ans da" for previous "McGurk" target).

Figure 4 - Results of Experiment 1: response times

(a) Mean response times for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (left) vs. incoherent (right) contexts.

(b) Variations of response times with context duration. For coherent vs. incoherent contexts and for "ba" vs. "McGurk" targets, the four successive bars display mean response times (and their standard errors) for 5-syllables, 10-syllables, 15-syllables and 20-syllables context durations from left to right.

Figure 5 - Stimuli setting for Experiment 2.

The coherent context is the same as in Experiment 1. The incoherent contexts are produced by editing the acoustic content of the coherent context, either shuffling syllables (incoherent P) or temporally shifting them (incoherent T) or applying both

modifications (incoherent PT). There is no discontinuity in the video track from context to target hence no transition stimulus is necessary.

Figure 6 - Results of Experiment 2

(a) Percentage of " ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (C) vs. phonetically incoherent (P), temporally incoherent (T) and phonetically and temporally incoherent (PT) contexts.

(b) Individual results for McGurk targets. For each of the 20 subjects, the percentage of "ba" responses for the McGurk target is displayed in the C, P, T and PT contexts.

(c) Effect of the precedent context, depending on the context of the present stimulus. Percentage of " ba " responses for McGurk stimuli in a coherent, incoherent P, incoherent T or incoherent PT context (from left to right), according to the precedent context (different drawing of individual bars).

(d) Analysis of responses to McGurk stimuli depending on context ("Coherent", P, T and PT), precedent target stimulus ("Prec Ba" vs "Prec McGurk") and previous answer ("Ans ba" for previous "Ba" target, "Ans ba" and "Ans da" for previous "McGurk" target).

Figure 7 - One-stage vs. two-stage model for audiovisual fusion in speech perception

(a) A possible one-stage model

(b) A possible two-stage model

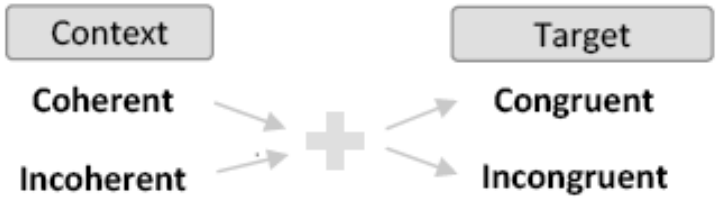
Context
Target
Coherent
Incoherent
Congruent
Incongruent

Figure 1

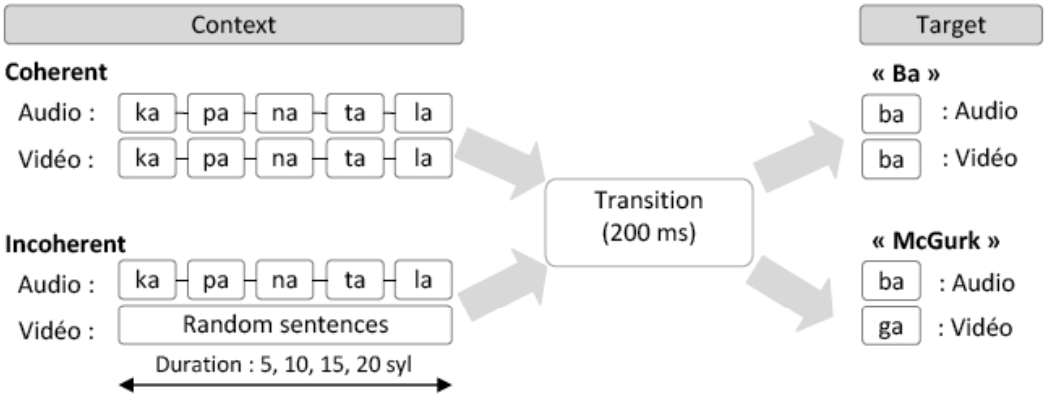
Context
Coherent
Audio :
ka
pa
na
ta
la
Vidéo :
ka
pa
na
ta
la
Incoherent
Audio :
ka
pa
na
ta
la
Vidéo :
Random sentences
Duration : 5, 10, 15, 20 syl
Transition
(200 ms)
Target
« Ba »
ba
: Audio
ba
: Vidéo
« McGurk »
ba
: Audio
ga
: Vidéo

Figure 2

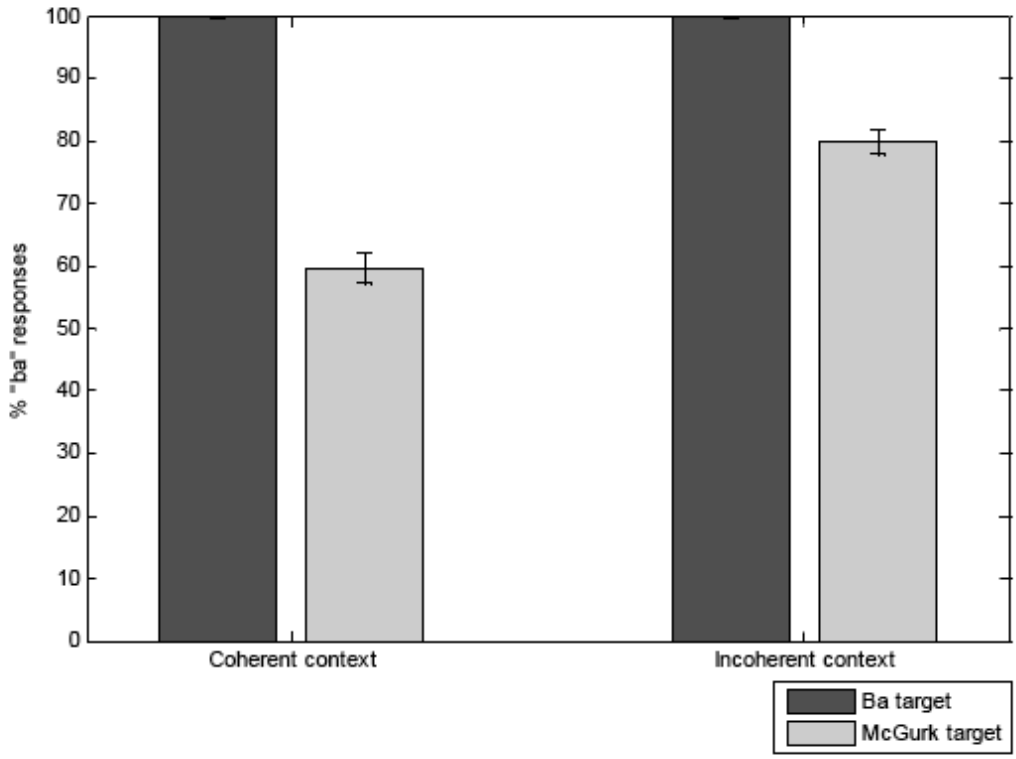
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Coherent context
Incoherent context
Ba target
McGurk target

Figure 3a

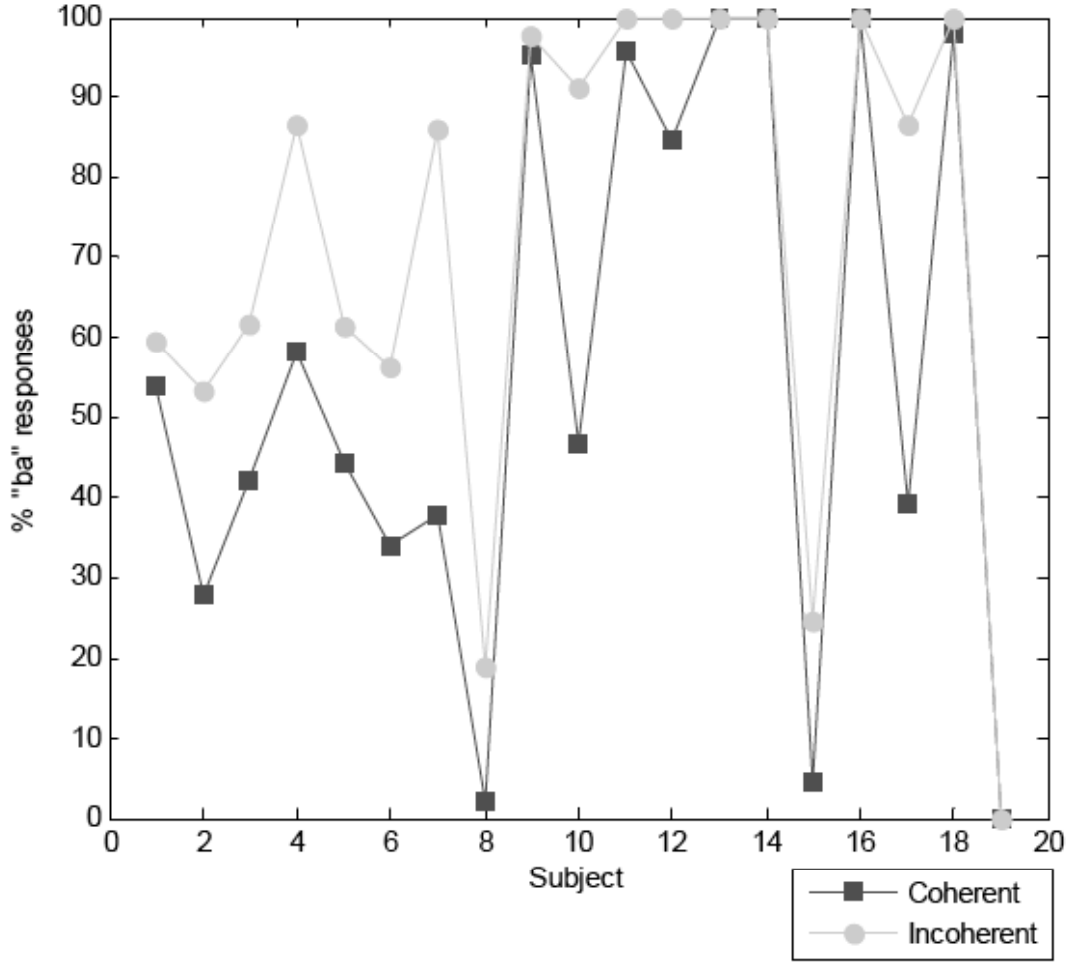
% "ba" responses
Subject
Coherent
Incoherent
0
10
20
30
40
50
60
70
80
90
100
0
2
4
6
8
10
12
14
16
18
20

Figure 3b

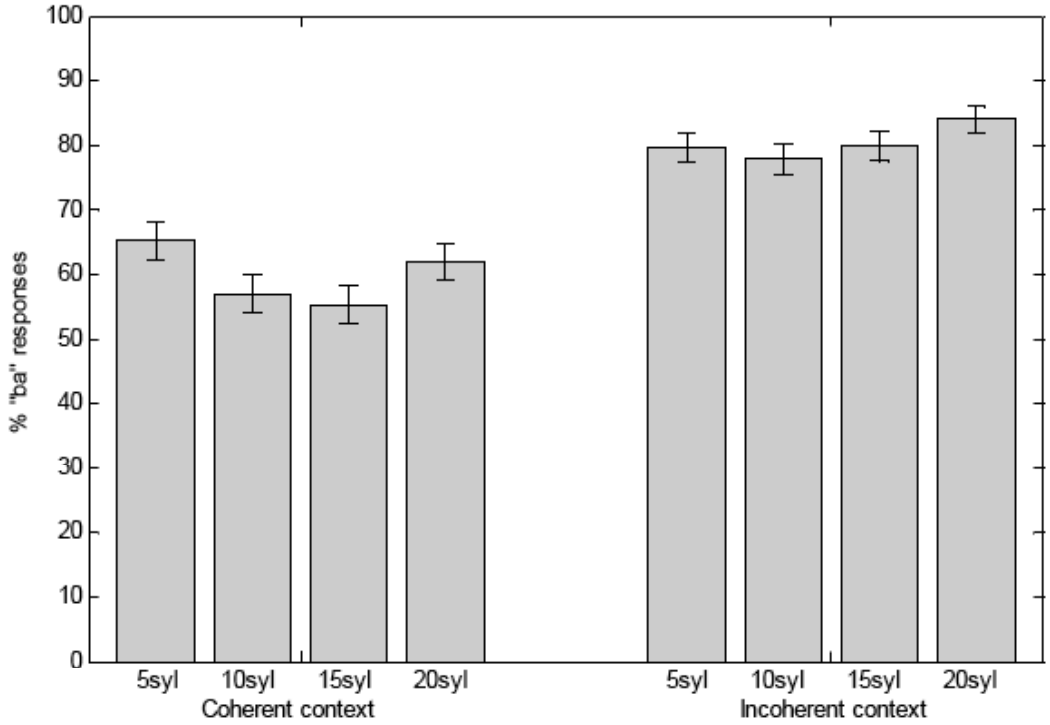
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
5syl
10syl
15syl
20syl
Coherent context
5syl
10syl
15syl
20syl
Incoherent context

Figure 3c

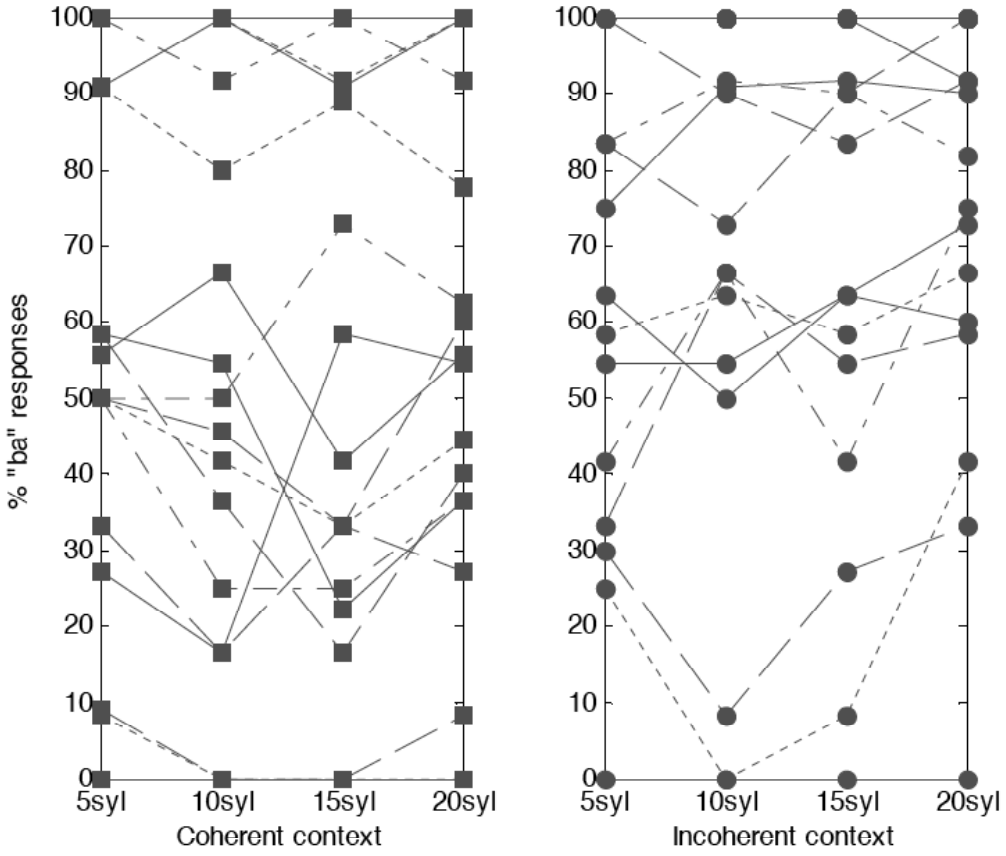
% "ba" responses
100
90
80
70
60
50
40
30
20
10
0
5syl
10syl
15syl
20syl
Coherent context
Incoherent context

Figure 3d

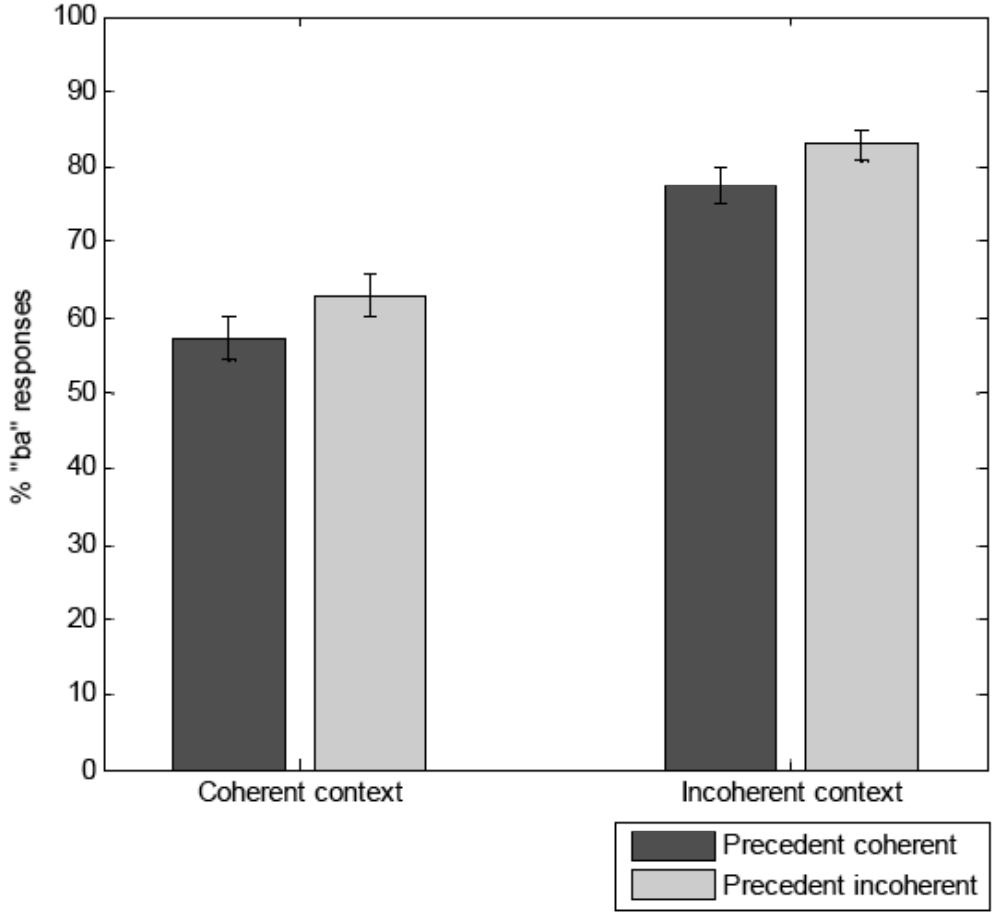
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Coherent context
Incoherent context
Precedent coherent
Precedent incoherent

Figure 3e

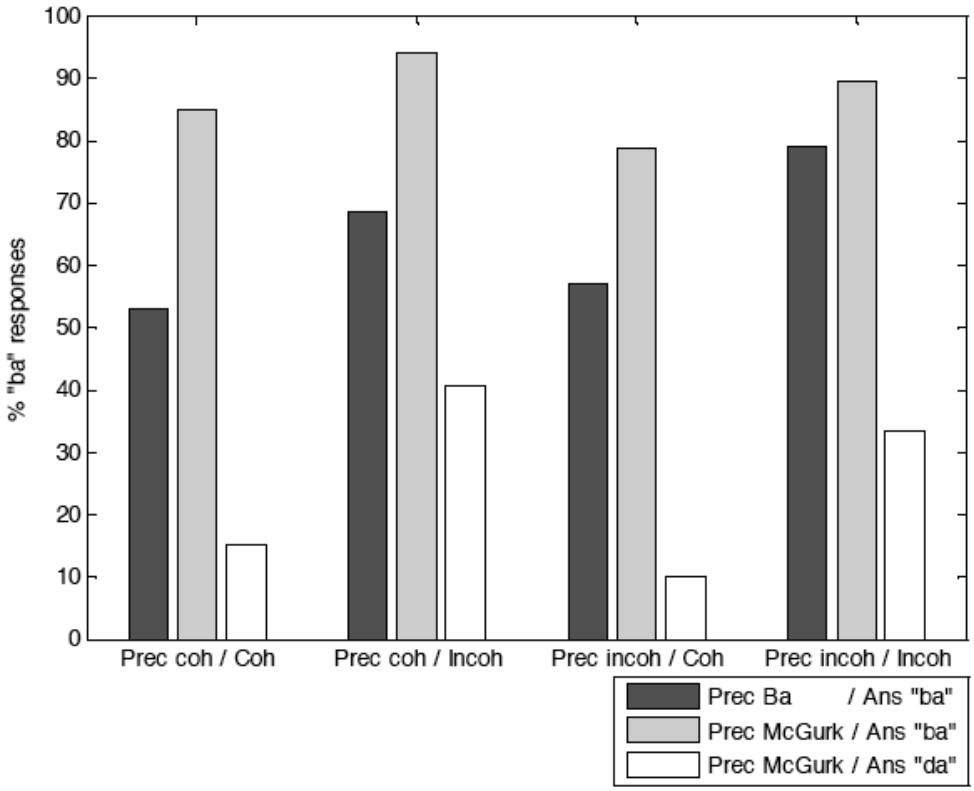
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Prec coh / Coh
Prec coh / Incoh
Prec incoh / Coh
Prec incoh / Incoh
Prec Ba / Ans "ba"
Prec McGurk / Ans "ba"
Prec McGurk / Ans "da"

Figure 3f

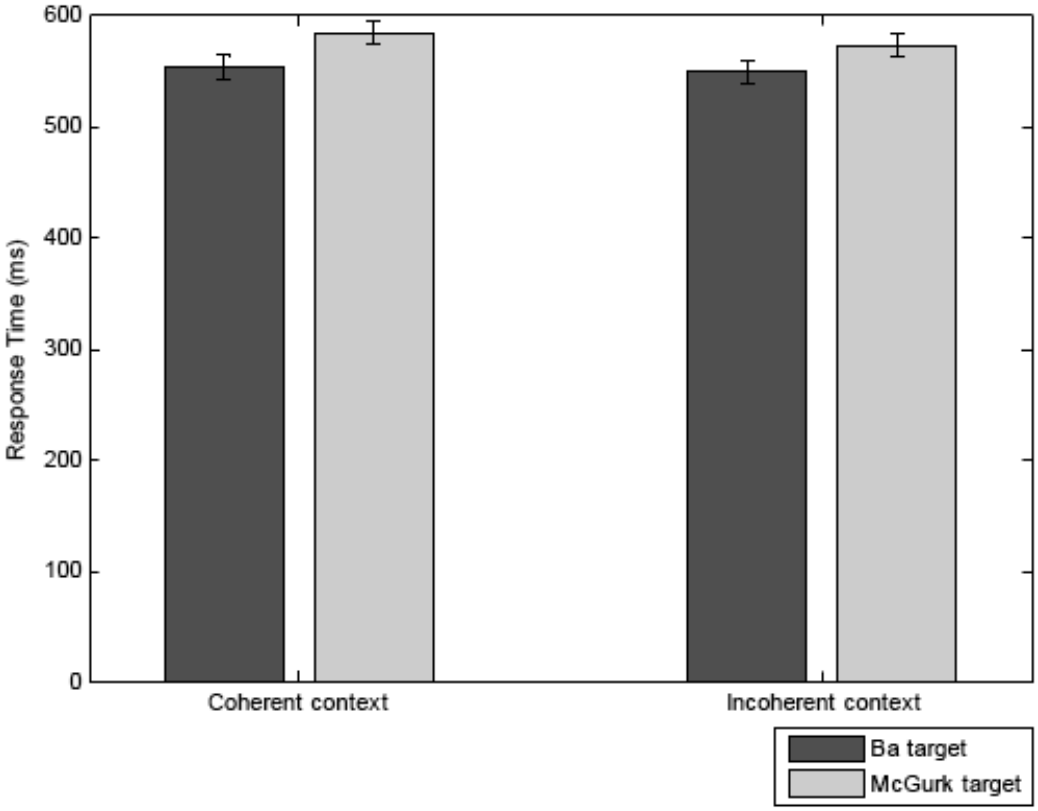
600
500
400
300
200
100
0
Response Time (ms)
Coherent context
Incoherent context
Ba target
McGurk target

Figure 4a

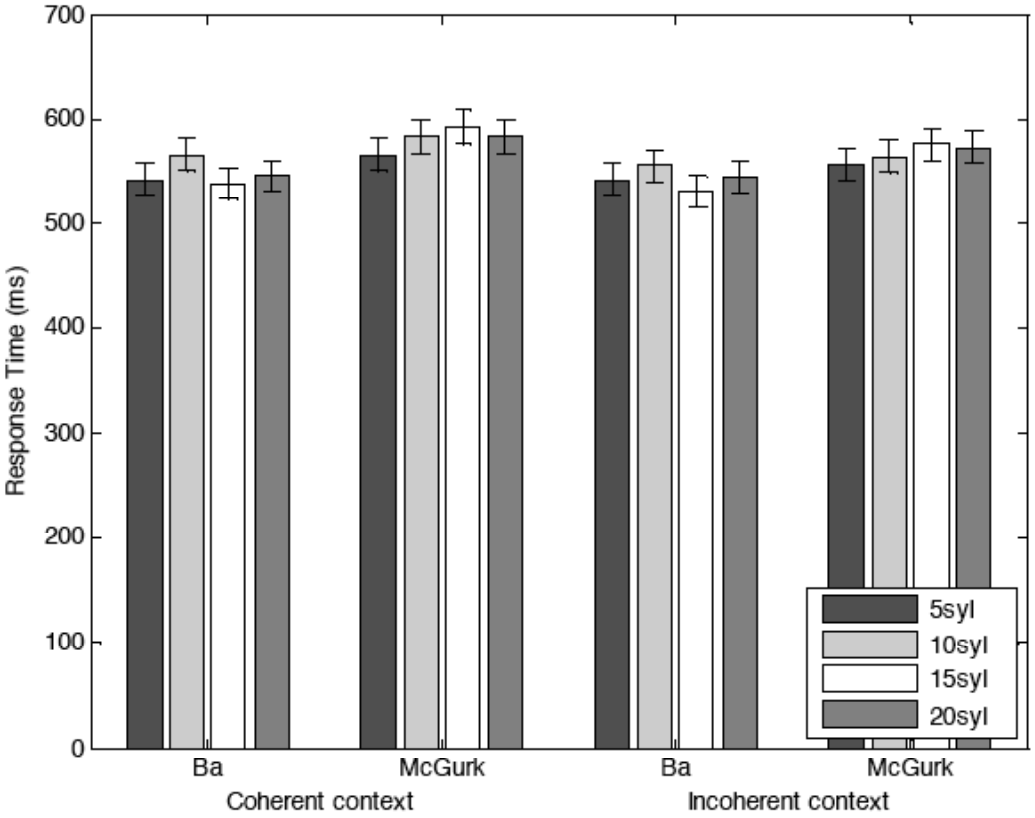
700
600
500
400
300
200
100
0
Response Time (ms)
Ba
McGurk
Ba
McGurk
Coherent context
Incoherent context
5syl
10syl
15syl
20syl

Figure 4b

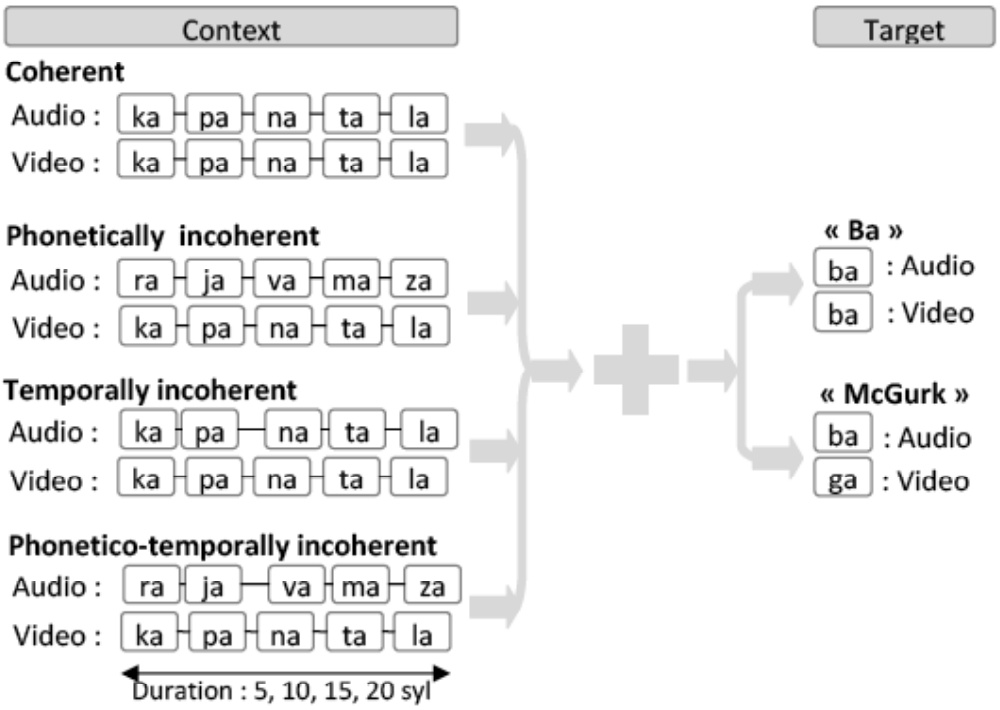
Context
Target
Coherent
Audio : ka pa na ta la
Video : ka pa na ta la
Phonetically incoherent
Audio : ra ja va ma za
Video : ka pa na ta la
Temporally incoherent
Audio : ka pa na ta la
Video : ka pa na ta la
Phonetico-temporally incoherent
Audio : ra ja va ma za
Video : ka pa na ta la
Duration : 5, 10, 15, 20 syl
« Ba »
ba : Audio
ba : Video
« McGurk »
ba : Audio
ga : Video

Figure 5

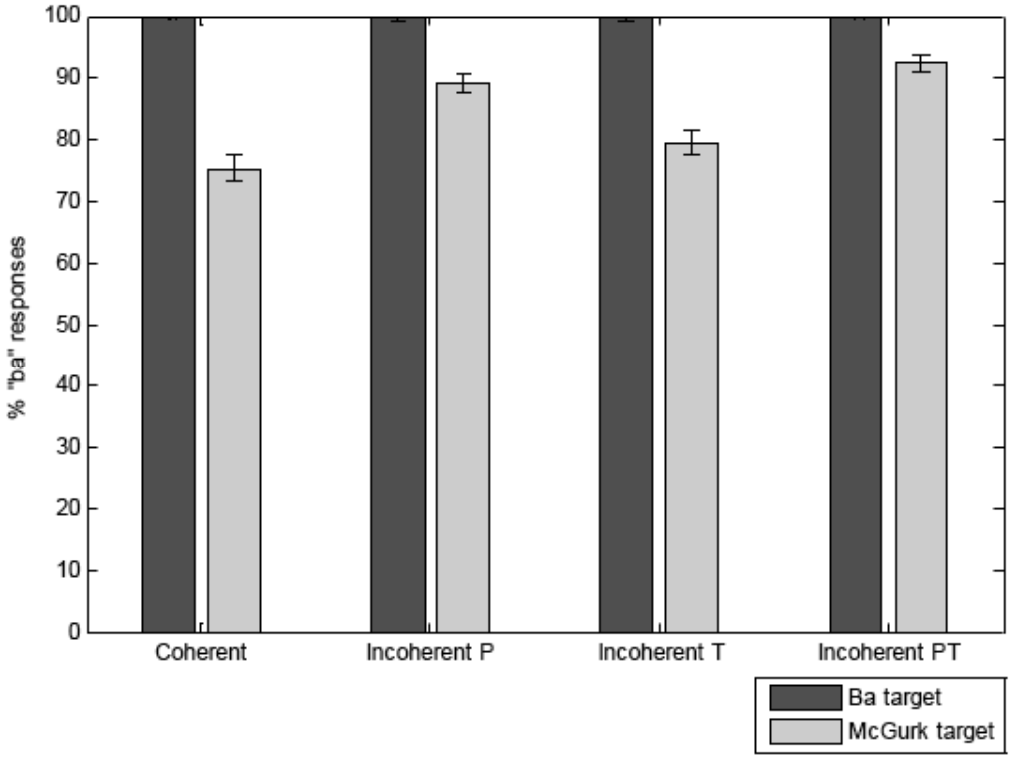
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Coherent
Incoherent P
Incoherent T
Incoherent PT
Ba target
McGurk target

Figure 6a

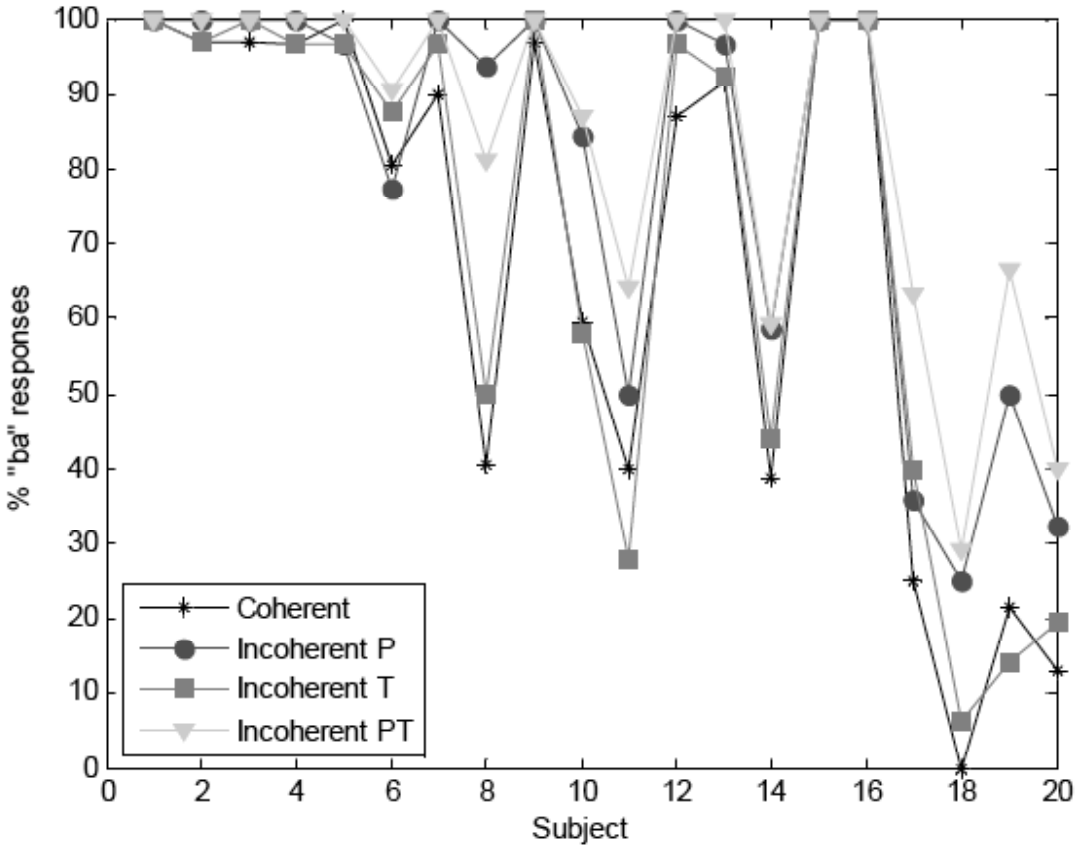
% "ba" responses
Subject
Coherent
Incoherent P
Incoherent T
Incoherent PT
0
10
20
30
40
50
60
70
80
90
100
0
2
4
6
8
10
12
14
16
18
20

Figure 6b

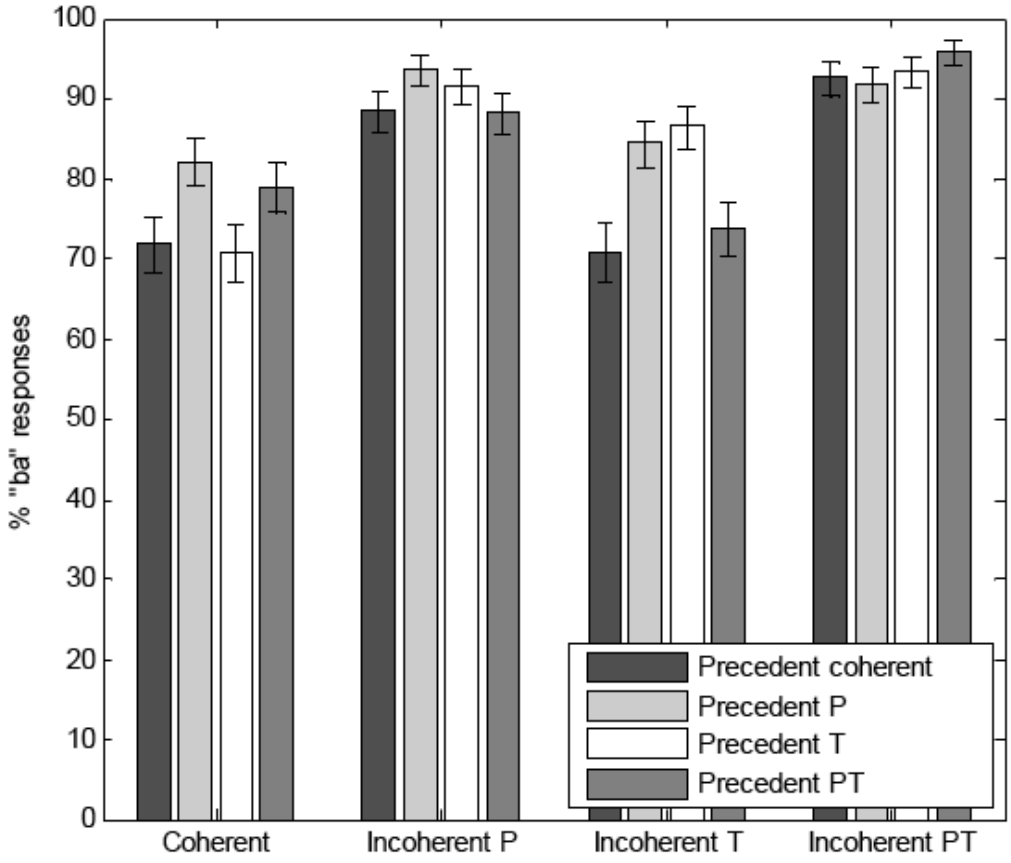
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Coherent
Incoherent P
Incoherent T
Incoherent PT
Precedent coherent
Precedent P
Precedent T
Precedent PT

Figure 6c

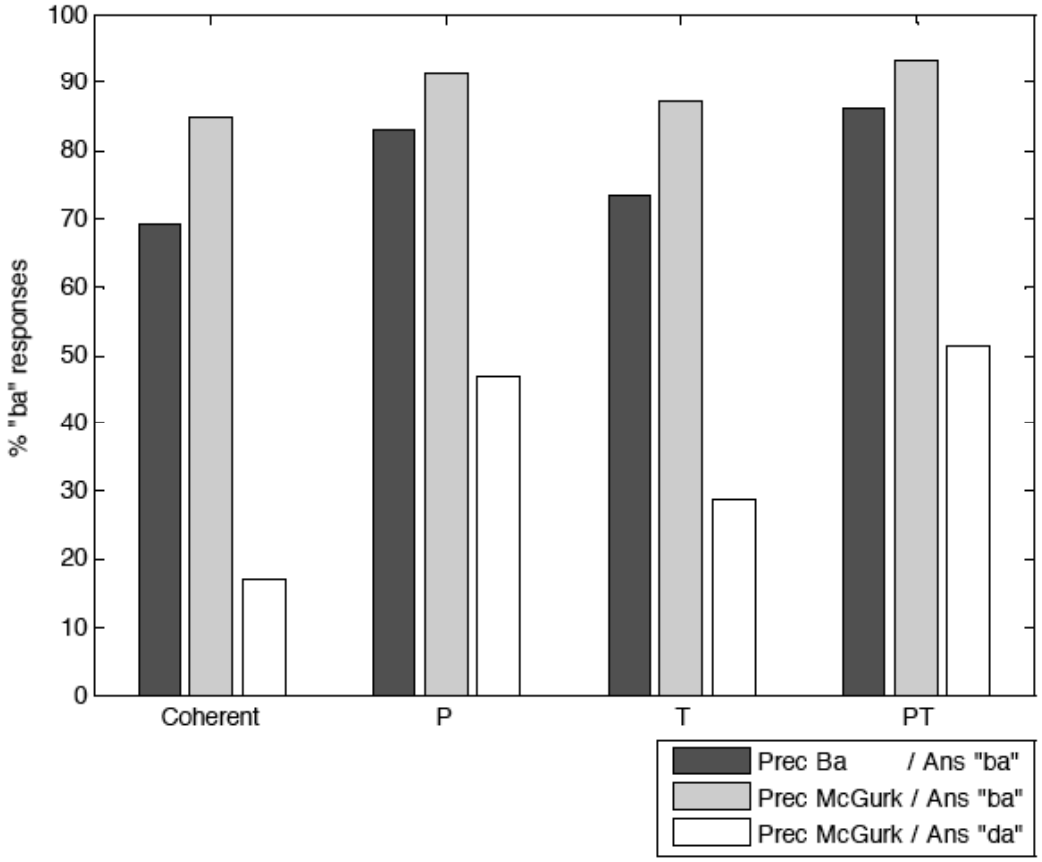
100
90
80
70
60
50
40
30
20
10
0
% "ba" responses
Coherent
P
T
PT
Prec Ba / Ans "ba"
Prec McGurk / Ans "ba"
Prec McGurk / Ans "da"

Figure 6d

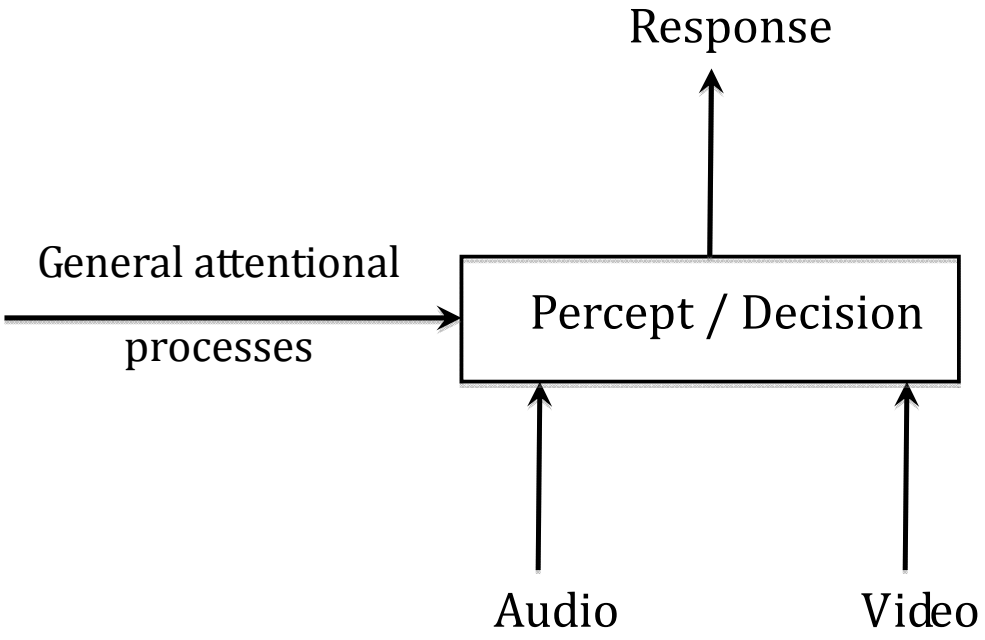

Figure 7a

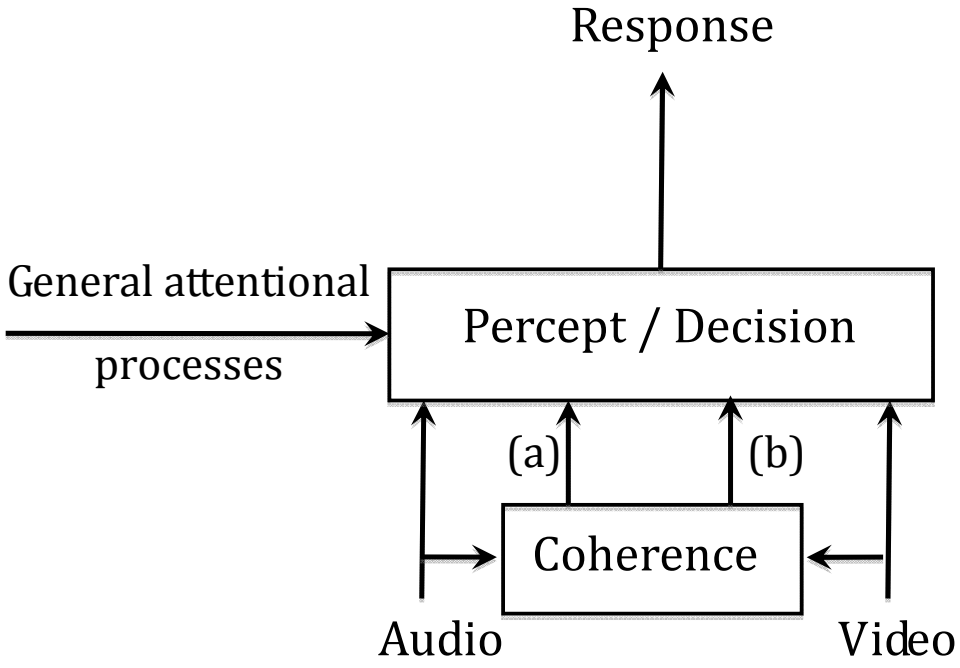

Figure 7b